A MONTE CARLO STUDY OF SEVERAL DIFFERENT APPROACHES TO THE BEHRENS–FISHER PROBLEM

Carlos R. Meléndez Román

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Education (Educational Psychology, Measurement and Evaluation).

Chapel Hill
2016

Approved by:

William B. Ware

Jeffrey A. Greene

Eric A. Houck

Mary R. Lynn

Todd A. Schwartz

ABSTRACT

Carlos R. Meléndez Román: A Monte Carlo Study of Several Different Approaches to the Behrens–Fisher Problem
(Under the direction of William B. Ware)

One of the tests most often used to compare the means difference of two independent groups is the pooled *t*-test (i.e., Student's *t*-test or classical *t*-test). However, the validity of pooled *t*-test results is based on certain assumptions, including the homogeneity of variance (HOV). The violation of HOV has been called in the statistical literature the Behrens–Fisher problem.

The purpose of this dissertation was to compare and contrast the Type I error-rate performance and statistical power of five solutions to the Behrens–Fisher problem under several different simulated conditions. The methods studied were the pooled *t*-test, the Cochran–Cox *t*-test, the *t*-test with the Welch–Satterthwaite correction, and two different bootstrap methods: a non-parametric bootstrapping method using the Efron and Tibshirani (1993) approach, and a non-parametric bootstrapping method using a modified version of the Good (2005) approach. These methods were compared and contrasted in terms of Type I error-rate performance and statistical power for several different conditions. In this study, computer simulated data from a normal population with mean 0 and variance 1 were used to contrast the achieved significance level (*p*-value) and power of the five method mentioned for testing the mean difference of two groups when the HOV could not be assumed.

Only three methods consistently yielded accurate *p*-values for a two-tailed hypothesis test: the Cochran–Cox *t*-test, the nonparametric bootstrap method using the Efron–Tibshirani approach, and the Welch–Satterthwaite approximate *t*-test. These methods effectively controlled for Type I error rate because the nominal and the empirical significance levels (α) were statistically equal.

Of these, only the Welch–Satterthwaite approximate *t*-test completely controlled the Type I error rates in all of the studied conditions. On the other hand, in almost all of the simulated conditions, the Welch–Satterthwaite approximate *t*-test was slightly more powerful than the other methods. However, in the special cases when the sample sizes were equal or when the variances were equal, the pooled *t*-test controlled the Type I error rates in almost all instances. Moreover, in those cases, the pooled *t*-test was also the most powerful method for detecting the mean differences, most of the time.

The present study presents no compelling evidence to indicate that a method other than the Welch–Satterthwaite approximate *t*-test provides a better alternative to the Behrens–Fisher problem, except when the sample sizes are equal. Given the evidence presented in the present study, of the five methods evaluated, I recommend use of the Welch–Satterthwaite approximate *t*-test in cases when the samples have been obtained from normally distributed populations, when the sample sizes are unequal, and when there is uncertainty that the variances of the samples are equal. In cases when the sample sizes are equal or when the variances are equal, I recommend use of the pooled *t*-test.

A todos los de mi familia que nunca han fallado en desearme lo mejor. En especial a mis hermanos (Carlos Alberto, Loaiza y Julia), a mi padre Benigno, a mi cuñada Yanira y a mis sobrinos (Yaphet y Carlos Yadiel). Gracias a todos ustedes por ser solidarios y apoyarme en todo lo que me he propuesto alcanzar.

A todos los de mi otra familia de Manatí y Morovis que siempre me tienen en sus pensamientos y oraciones. Especialmente a Margarita y a Pablo porque durante todos estos años han estado muy pendientes de mí y su apoyo y ayuda han sido muy significativos. ¡Muchas gracias!

No puedo dejar de dedicar esta disertación también a la memoria de tres personas que fueron muy especiales en mi vida. Estas fueron, mi madre Carmen Luz, mi madrina Juana y mi tía María Antonia. Las enseñanzas e influencias obtenidas de cada una de ellas me acompañarán toda mi vida. Siempre estaré muy agradecido por todo el cariño, apoyo y ayuda que de diversas formas recibí de las tres, al igual que por la oportunidad que tuve de compartir distintas etapas de este tramo con cada una de ellas y porque cada una me regaló parte de su sabiduría.

Finalmente, pero muy importante, a José R. Vázquez, cuya amistad supera por mucho todo lo imaginable y cualquier expectativa. Gracias por haberte convertido en mi consejero, sostén, confidente y escuchador, entre los muchos roles que has desempeñado por años. Gracias también por haber sido mi ayudante y cómplice en esta carrera que felizmente culmina con esta disertación. *¡Finalmente, lo logramos!*

¡A todos, mis más sinceras gracias!

ACKNOWLEDGEMENTS
(Agradecimientos)

Primeramente, gracias a todos los miembros de mi Comité de Disertación. Algunos han estado conmigo desde el comienzo, como parte de mi Comité de Programa de Estudios. Otros dijeron presente, dispuestos a cooperar, en el momento preciso. ¡Muchas gracias a todos! Fue un honor para mí el que ustedes fueran los que supervisaran y posteriormente aprobaran mi disertación. Muchas gracias Dr. Greene por su tiempo y por dar la milla extra al velar con celo para que mi trabajo fuera de la mejor calidad. Sus revisiones cuidadosas a las diferentes versiones del manuscrito, junto a sus observaciones y comentarios, me provocaron para que produjera una disertación aún mejor de lo anticipado. Siempre estaré muy agradecido con ustedes, doctores Houck y Lynn, por su tiempo y cooperación, aceptando el integrarse a mi Comité de inmediato, tan pronto los necesité, para aportar experiencia, conocimientos y total apoyo. Muchas gracias Dr. Schwartz por aportar sus conocimientos y por sus revisiones del manuscrito. Igualmente, gracias por sus preguntas estimulando mi razonamiento, buscando y logrando la mejor calidad, tanto en la investigación como en el manuscrito de esta disertación. Además gracias por haber estado disponible para ayudarme a obtener la concentración menor en Bioestadística. Por otro lado, gracias por su tiempo y por todos sus consejos académicos, por todas sus palabras de estímulo, por estar siempre pendiente y disponible para ayudarme al igual que por convertirse en otro de mis consejeros sobre la carrera profesional que deseo seguir. Quiero agradecer también al Dr. Daniel J. Bauer por su tiempo y por haber formado parte del Comité de Programa de Estudios.

Muchas gracias por igual a los profesores de la Facultad de Educación (“School of Education”) de la University of North Carolina en Chapel Hill, particularmente a los del programa de “Educational Psychology, Measurement, and Evaluation”. Gracias por toda su comprensión y todo su apoyo durante estos años en que he sido estudiante de su facultad.

Por otro lado, deseo agradecer a algunos de los otros profesores que he tenido durante el transcurso de mi vida como estudiante universitario. Particularmente al Dr. Brian Neelon y a la profesora Emma Borynski. Gracias a las recomendaciones de ustedes tuve la oportunidad de ser admitido a este programa de doctorado. Les estaré muy agradecido por confiar en mí y por haberme apoyado. Igualmente le agradezco al Dr. Roberto E. Torres Zeno por haberme apoyado con una de las mejores cartas de recomendación que he podido leer en mi vida. Mención aparte merece el agradecimiento y aprecio que le tengo al Dr. Josué Guzmán. Gracias por desearme siempre lo mejor, por sus consejos, apoyo, por ser un mentor, por estimularme para comenzar y completar el doctorado y por confiar en mí.

Deseo expresar también mi gratitud al Dr. Sandip Sinharay por haber sido el mejor mentor de internado en investigación que cualquier estudiante desearía tener. Gracias por haberme dado la oportunidad de publicar un artículo junto a ti. Igualmente gracias por continuar brindándome tu apoyo, por confiar en mí y por tus cartas de recomendación. Gracias a los doctores, Bradley Efron y Fortunato Pesarin por haber sacado de su tiempo para responder al mensaje de correo electrónico que le envié a cada uno por separado con preguntas sobre el tema de esta disertación y por la valiosa información que me brindaron.

He querido dejar para el final mi agradecimiento especial al Dr. William B. Ware, mi “advisor”. Sin embargo, no lo he dejado para el final porque sea menos importante que los demás. Por el contrario, todas las actitudes y cualidades por las que estoy agradecido de los que

he mencionado anteriormente, servirían también para describir perfectamente al muy estimado Dr. Ware. A él le estaré eternamente agradecido por haber sido mi mentor principal, por todo el tiempo que me ha dedicado, por su apoyo genuino, por sus recomendaciones, por estar siempre pendiente de mi bienestar, por su estímulo para que yo continuara luchando y alcanzara mis metas, por lograr la mejor calidad de esta disertación y por su paciencia. Yo podría continuar describiendo las cualidades del Dr. Ware y explicando las razones por las cuales le estoy muy agradecido y le guardo tanto aprecio. No obstante creo que la siguiente oración resume todo lo que significa para mí el haber tenido el honor de que Dr. Ware haya sido mi consejero académico y director tanto de mi Comité de Programa de Estudios como de mi Comité de Disertación. **Siempre estaré en deuda de gratitud con usted, Dr. Ware, porque sin su total apoyo, tiempo, comprensión, ayuda y paciencia, yo no hubiese podido terminar esta disertación y mucho menos completar mis estudios de doctorado**. **¡Muchas gracias por todo!**

Por último, deseo indicar que sería imposible mencionar a todos los que se merecen estar incluidos en esta sección de agradecimientos. El espacio disponible es insuficiente. De todas formas, muchas gracias a todos los que de alguna forma me apoyaron o colaboraron conmigo para que yo pudiera lograr esta meta.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction and Background

Introduction

In this study, the problem of comparing the means of two groups under the uncertainty of equal sample variances was examined. According to Alba Fernández, Jiménez Gamero, & Muñoz García (2008), "the problem of testing whether two samples come from the same or different populations is a classical one in statistics" (p. 3731). This problem is a special case of comparing two or more groups (van Belle, Fisher, Heagerty, & Lumley, 2004). Which statistical test should be employed for the comparison in a particular study depends on several factors, including the number of groups, the distribution of the variable(s) of interest within the population (e.g., mean and variance), the purpose of the comparison, the type of sampling employed to obtain the data, and the known information the statistician may have about the characteristics of the available data (e.g., mean and variance of the sample).

According to Kim and Cohen (1998), testing the difference between the means of two populations (i.e., two groups) is a very common task for statisticians. One of the tests most often used to compare the means difference of two independent groups (i.e., two independent samples) is the pooled *t*-test based on the *t* statistic (i.e., Student's *t*-test or classical *t*-test). This parametric test is very popular among statisticians because it is both relatively simple to compute and readily available in every common statistical computer package. Additionally, the pooled sample *t*-test requires only knowing the two sample sizes as well as two types of easily obtained statistics: the sample means and the sample standard deviations.

However, the validity of *t*-test results is based on certain assumptions. First of all, parametric tests, including the *t*-test, "rely on a mathematically known but assumption-constrained sampling distribution to derive probabilities" (Hayes, 2000, p. 653). One *t*-test assumption about the sampling distribution is that the population is normally distributed with mean μ and common variance σ^2 (Howell, 2002; Moore & McCabe, 2004; van Belle, Fisher, Heagerty, & Lumley, 2004). Another important assumption of this test is the independence of observations (i.e., that the error component of any observation is unrelated to the error component of any other observation) (Moore & McCabe, 2004).

The assumption of common variance σ^2 is often characterized as the assumption of homogeneity of variance (HOV). This assumption, which essentially states that the two sample variances (s_1^2, s_2^2) are estimating the same population variance or σ^2 (Moore & McCabe, 2004), allows the statistician to combine or pool the sample variances to form an estimate of the common population variance (Howell, 2002; van Belle, Fisher, Heagerty, & Lumley, 2004). A violation of the HOV assumption carries serious consequences for *t*-test results because the *t* statistic can no longer produce valid or accurate[1] results (Welch, 1938).

It is generally accepted that the *t*-test is the uniformly most powerful unbiased (UMPU) test for the equality of two independent population means, if the assumptions of normality and HOV are not violated (Olejnik & Luh, 1994). However, if the variances are heterogeneous, a condition known as heteroscedasticity, the Type I error rate (i.e., the significance level of the test) is no longer stable (Olejnik & Luh, 1994). That is, the significance level of the *t*-test could actually be greater or lower than the pre-assigned level (i.e., the significance level set by the

[1] In this dissertation, a method was considered that yield accurate *p*-values if it effectively control for Type I error rate. In other words, when the nominal and the empirical significance levels (α) were statistically equal the method was considered as one that yield accurate *p*-values.

researcher; nominal α). Results obtained with this test may not be valid or accurate if all assumptions of the test, including the HOV, are not met. This situation has been called in the statistical literature the Behrens–Fisher problem (Howell, 2002; Kim & Cohen, 1998; Pesarin, 1995). In another section of this chapter, I shall describe this problem in more detail.

Purpose of the Study

The main purpose of the present study was to compare and contrast the Type I error-rate performance and statistical power of five approaches (tests or methods) to the Behrens–Fisher problem under several different simulated conditions. The methods studied were the UMPU test (i.e., *t*-test or classical *t*-test), the Cochran–Cox *t*-test assuming unequal variances, the approximate *t*-test with the Welch–Satterthwaite correction, and two different bootstrap methods that are described in detail in chapters 2 and 3. These approaches to the Behrens–Fisher problem were compared and contrasted in terms of Type I error-rate performance and statistical power, given several different conditions.

Illustration

A statistical exercise presented by Green and Salkind (2005, pp.173–174) illustrates the problem with the violation of the HOV assumption and its implications with respect to statistical analysis using the *t*-test. The purpose of that exercise was to determine whether the inclusion (i.e., integration) of seventh-graders receiving special education (i.e., independent variable) in regular-instruction classes academically hurt or helped the children who were receiving regular instruction. In other words, the problem called for an assessment of the effectiveness of integration on academic achievement for the children in regular education. The achievement comparison (dependent variable) of these seventh-graders was based on the difference between a standardized test administered at the beginning of the year (pretest) and at the end of the year

(posttest). The data of this exercise consisted of two independent groups, (integrated and non-integrated classrooms[2]) and one dependent variable (change in achievement calculated as the difference between posttest and pretest scores).

Table 1 contains descriptive statistics of the data, which were obtained using SAS software[3]. These data contain some interesting characteristics. The first, obviously, is the inequality of the sample sizes. Another is that the mean difference in change scores (M_d = 8.07), which is based on average change of the integrated group (M_i = 9.6) and the average change of the not-integrated group (M_n = 1.53), seems relatively large; this difference suggests a difference in the achievement of the two groups. A third characteristic is that skewness and kurtosis values (as well as a visual inspection of the histograms and of the normal probability plot or QQ plot, which are not presented here) do not reveal any apparent problem in terms of a possible violation of the normality assumption—a necessary characteristic for parametric tests such as the *t*-test.

Table 1

Descriptive Statistics for Green and Salkind (2005)

Dependent Variable	Independent Variable (classrooms)	*N*	Statistic			
			Mean	Std. Dev.	Skewness	Kurtosis
Post–Pre	Integrated	25	9.60	16.54	.62	.270
	Non-integrated	15	1.53	7.41	-.71	-.002

Moreover, the results of the statistical test of normality (Shapiro & Wilk, 1965) do not show statistically significant results for any group (W_i = .94, p =.13; W_n = .95, p =.55). These

[2] The term "integrated classrooms" refers to those classrooms into which special-education students have been integrated for regular instruction; "non-integrated classrooms" refers to classrooms that contain no special-education students.

[3] SAS software (Versions 9.2) was used for all analyses of this section.

results provide additional evidence in favor of the normality assumption. Given that the skewness and kurtosis results do not suggest any major problem with that assumption, and taking into account that the Shapiro–Wilk results are not statistically significant, it should be concluded that the data of both groups appear to be relatively normally distributed.

Because the two groups (integrated and non-integrated classrooms) were independent and the dependent variable (posttest–pretest) was both continuous and normally distributed, one might think that a two-independent-sample *t*-test can be used for evaluation if the difference between the means of the two groups is statistically significant (Green & Salkind, 2005). Therefore, I decided to conduct the *t*-test analysis for that exercise. Before evaluating the results of *t*-test statistics, however, the HOV should be evaluated to verify if a violation of the equal variance assumption has occurred.

Table 2 contains the results of two of the statistical procedures developed for testing the equality of the variance. These are Levene's test (1960), also called the homogeneity of variance test, and the folded form of the *F* statistic, *F*'. Both test the hypothesis that the variances are equal.

Table 2

Homogeneity (Equality) of the Variance Tests

Test	Statistic	*p*-value
Levene's	$F = 4.10$	.050
Folded F	$F' = 4.99$	.003

Given the significance results of the two test statistics evaluated for HOV at a significance level of .05, there is evidence to reject the null hypothesis that the variances are equal (i.e., the two samples do not estimated a common variance). This conclusion presents a problem for the analysis of the data, however, in terms of both the selection of an appropriate

statistical test and the validity of the results. A significant result on the HOV tests suggests that the *p*-value of the classical *t*-test result is most likely not accurate (Green & Salkind, 2005). Specifically, "if there are differences in variance combined with differences in sample size, the *t*-test is conservative when the larger group has the larger variance, and liberal when the smaller group has the larger variance" (Hayes, 2000, p. 655).

In the Green and Salkind (2005) exercise, if the significance result of HOV was ignored, the *t*-test result, assuming equal variances, would be $t(38) = 1.78$, $p = .08$. Based on this result, the conclusion would be that inclusion had a non-significant effect on the achievement test scores of regular-education students at a significance level of .05. However, if the significance of the HOV tests is taken into consideration, the Welch-corrected result, not assuming equal variances, would be $t(35.85) = 2.11$, $p = .04$. Based on this second result, the conclusion would be that achievement test scores of regular-education students in integrated classrooms improved more than those of regular students in non-integrated classrooms at a significance level of .05.

Those two *p*-values, one assuming and the other not assuming equal variances, are clearly contradictory because they suggest two different conclusions. Obviously, the contradiction of those results may be the result of the violation of HOV assumption, because the other two-independent sample *t*-test assumptions were met.

The Behrens–Fisher Problem

As previously mentioned, the violation of the unequal variances assumption when comparing the difference of two independent means, as illustrated by the Green and Salkind (2005) exercise, has been called in the statistical literature the Behrens–Fisher problem (Howell, 2002; Kim & Cohen, 1998; Pesarin, 1995) in reference to the first two statisticians who are known to have dealt with it. According to Dudewicz, Ma, Mai, & Su (2007), this problem "dates

back to the early twentieth- century work of astronomer Behrens in 1929 and the statistician Fisher in 1935" (p. 1584).

Pesarin (1995), a prolific researcher of the problems of statistical testing in the presence of heteroscedasticity (i.e., unequal variances), described the Behrens–Fisher problem as one that "regards the equality of the mean values of two normal distributions when variances are unknown and possibly unequal" (p. 131). The Behrens–Fisher problem occurs when it is necessary "to test the null hypothesis that the locations, but not necessarily the variances, are equal" (Neuhäuser, Lösch, & Jöckel, 2007, p. 5057). Kim and Cohen (1998) mentioned that this "problem arises when one seeks to make inferences about the means of two normal populations without assuming either that the variances are equal or that the ratio of variances is known" (p. 356). In other words, this "problem concerns the inference for the difference between the means of two normal populations whose ratio of variances is unknown" (Ghosh & Kim, 2001, p. 5) or is not equal to 1 (Sawilowsky, 2002). Other authors (Dudewicz et al., 2007; Scheffé, 1970; Wang & Chow, 2002) have published similar descriptions of the Behrens–Fisher problem.

As these descriptions have stated, the main issue is with the assumption of equality (i.e., homogeneity) of unknown variances (Hyslop & Lupinacci, 2003). Although this issue could also arise in hypothesis testing of the equality of the means of more than two samples, the present study is concerned only with the two-sample univariate case of this problem, as per the data from the preceding *t*-test example.

As previously stated, the two-independent-samples *t*-test is the classical parametric test used when the statistician is interested in obtaining the probability of the statistical difference between the means of two independent samples in the case of unknown, but presumably equal, variances. It is known that parametric methods are more powerful that nonparametric methods,

but only if the assumptions (e.g., normality and HOV) of the former methods are valid (Stonehouse & Forrester, 1998). However, "if these assumptions do not hold, then the parametric tests are considered to be less robust than nonparametric tests, i.e. more likely to report the null hypothesis to be false when, in fact, it is true (a Type I error)" (Stonehouse & Forrester, 1998, p. 63).

A Recommended Approach

Howell (2002) provided a relatively complete description of some guidelines that should be followed for the consideration of two assumptions and one side condition before conducting a *t*-test, when the goal is meaningful interpretation of the results. As mentioned on p. 3, the first assumption is the normality of the dependent variable and the second is homogeneity or equality of variance (HOV). The side condition, however, is related to equal versus unequal sample sizes. In the next few paragraphs I shall discuss the implications of violating each *t*-test assumption.

There is a vast literature that includes sufficient research and evidence to assume that the independent *t*-test is robust with regard to the violation of normality. In other words, the result of that test is not greatly affected if the data contain a moderate departure from this assumption (Howell, 2002; Stonehouse & Forrester, 1998). However, if "the distributions are markedly skewed (especially in opposite directions), serious problems arise unless the variances are fairly equal" (Howell, 2002, p. 215). Stonehouse and Forrester (1998) similarly remarked upon radical departure from normality and the robustness of the *t*-test.

However, departure from the HOV assumption has serious implications for a *t*-test result. The effect of the violation of the HOV assumption on a *t*-test result is closely related to the side condition concerning sample sizes. For equal sample size, HOV violation does not greatly affect whether results can be obtained (Stonehouse & Forrester, 1998). Instead, the main problem with

the violation of the HOV is that if the samples sizes are not equal, "the results are more difficult to interpret" (Howell, 2002, p. 215). According to Boneau (1960), if unequal sample sizes and unequal variance are combined, this combination "automatically produces inaccurate probability statements which can be quite different from the nominal values" (p. 62). The interaction between unequal sample sizes and variances in a *t*-test has also been emphasized by other authors, such as Stonehouse and Forrester (1998) and Glass, Peckham, and Sanders (1972).

It is important to keep this relatedness between HOV violation and sample sizes in mind because statisticians do not always have enough evidence to reasonably conclude that the assumptions of a particular parametric statistical test have been met. For example, in the exercise from Green and Salkind (2005), not only was the HOV assumption unmet but also, and even worse, the sample sizes were different and could be considered small. I believe that this is one of the most extreme situations, in terms of assumptions, for a parametric *t*-test analysis. This type of situation is frequently found in social science studies (e.g., education), in which controlled experiments are obviously less common than in the so-called "hard" sciences.

The Behrens–Fisher problem "has received considerable attention for decades" (Ghosh & Kim, 2001, p. 5). Nonetheless, van Belle et al. (2004) emphasized that it is only "of theoretical interest in statistics because there is no exact solution to such an apparently simple problem" (p. 139). Kim and Cohen (1998) made a similar observation. Hayes (2000) also noted the absence of "perfect solutions to the difficulties produced by variance heterogeneity when comparing group means" (p. 655). In other words, there is neither a direct theoretical solution to the Behrens–Fisher problem nor a uniformly most-powerful unbiased (UMPU) test for all sample sizes (Heiser, 2006).

In summary, when the HOV assumption is violated, the validity of the *t*-test is questioned. Still, neither a direct theoretical solution to the Behrens–Fisher problem nor a UMPU test for all sample sizes has been found, although several methods (tests or approaches) have been suggested and studied.

In Chapter 2, I review several solutions to the Behrens–Fisher problem that have been proposed, studied, and discussed. I shall also present a literature review of alternatives or approaches to this problem that have been proposed by statistical researchers. In Chapter 3, I present the research design of the current study, the purpose of which is to analyze the Type I error rates and power of several approaches or methods proposed in the literature as statistical tests to resolve the Behrens–Fisher problem (i.e., to obtain an accurate *p*-value for hypothesis testing when the HOV assumption is violated). The five approaches or methods considered herein are the Cochran and Cox *t*-test assuming unequal variances, a nonparametric bootstrapping method using the Efron and Tibshirani (1993) approach, a nonparametric bootstrapping method using a modified version of the Good (2005) approach, a pooled *t*-test or classical *t*-test, and an approximate *t*-test with a Welch–Satterthwaite correction.

In Chapter 4, the results of the present study, including comparisons of Type I error rates and power analysis properties among the five approaches or methods, are examined for different sample sizes and variances ratios using simulated data via a Monte Carlo study. The final chapter of this study (i.e., Chapter 5) contains a discussion about the results and their implications. Chapter 5 also contains sections about the limitations and problems encountered during this study as well as conclusions derived from the present study and recommendations for future practice.

Chapter 2: Review of the Literature

Numerous methods or approaches have been proposed to resolve the Behrens–Fisher problem (Aspin & Welch, 1949; Efron & Tibshirani, 1993; Fisher, 1935; Hyslop & Lupinacci, 2003; Lee & Gurland, 1975; Satterthwaite, 1946; Sawilowsky, 2002; Scheffé, 1970; Wang & Chow, 2002; Welch, 1947). According to Scheffé (1970), in the case of Behrens–Fisher problem, the "normality assumption is of no practical importance for any of the solutions considered [because these] are robust against its violation" (p. 1501). That is, the validity of the results of the solutions that have been tried for the violation of HOV assumption is not affected by a concomitant violation of the normality assumption. Therefore, most of the solutions are essentially sets of procedures that deal with the problem of heteroscedasticity.

The Behrens–Fisher Solution

The first to offer a solution to this problem was Behrens (1929); later, the solutions was reframed by Fisher (1935), who justified Behrens's solution by using the fiducial theory of inference (Ghosh & Kim, 2001; Sawilowsky, 2002). In fact, Fisher was convinced that he had found "the exact test for the significance of the difference, *d*, between the observed means, equivalent to that given in 1929 by W. -V. Behrens" (Fisher, 1935, p. 397). Therefore, the Behrens–Fisher problem can also be described as the problem of finding the distribution of this particular approximation of the *t* statistic (van Belle et al., 2004).

Other Parametric Solutions

Another parametric alternative to the Behrens–Fisher problem is the use of an approximation of the *t* statistic instead of the actual statistic. The problem with such an

approximation, however, is that its distribution is not the same as the distribution of the *t* statistic. Still, there are ways to apply this approximation. According to Wang and Chow (2002), the two most common practical approximate methods are the Cochran and Cox and the Satterthwaite, although the most commonly found parametrical alternative in today's statistical software packages is the Welch–Satterthwaite solution. However, the statistical computer software that I used for analysis in the present study can easily supply the Cochran and Cox *t*-test p-value.

It should be noted that, according to Wang and Chow (2002), the Cochran–Cox and the Satterthwaite methods perform comparably well if the data are normally distributed; however, the latter performs somewhat better if there is a violation of the normality assumption. By contrast, Heiser (2006) mentioned that although the Welch–Aspin–Satterthwaite solution (another name for the Welch–Satterthwaite) is a solution to the Behrens–Fisher problem, it is not robust to the violation of normality. Moreover, Lee and Gurland (1975) found, after conducting size (i.e., significance level or α) and power tests of various solutions to the Behrens–Fisher problem, that although the Cochran and Cox solution is simple (and was one of the most widely used at that time), the Type I error rate of this test tends to be conservative; this is not a problem with large samples.

The solution originally proposed by Smith in 1936, which is the same as the solution published by Satterthwaite in 1941 and 1946, is equivalent to the Welch approximate *t*-test solution (Davenport & Webster, 1975). This equivalency explains why this approximate *t*-test solution is commonly known as Welch–Satterthwaite solution. Both Welch (1938) and Satterthwaite (1946) emphasized that when there is no *a priori* evidence that the ratio of the variances is equal to 1 (i.e., variances are equal), it is more appropriate to use a test that

approximates the *t*-statistic instead of directly using the *t*-statistic to compare the equality of two independent means. The Welch–Satterthwaite approximation uses a different estimate for each variance instead of applying the pooled variance to compute the *t*-statistic. Thus the main contribution of these two researchers consisted of a solution for the computation of the adjusted degrees of freedom, in order to find the correct or at least compute a more accurate *p*-value for the *t*-test result when the approximation of the *t*-statistic is used (Satterthwaite, 1946; Welch, 1938).

According to Scheffé (1970), the "Welch's approximate *t*-solution, which requires only the ubiquitous *t*-tables, is a satisfactory practical solution of the Behrens-Fisher problem" (p. 1505). Similarly, Davenport and Webster (1975) stated that the Welch's approximate *t*-test is "one of the most practical solutions [to the Behrens-Fisher problem] (and by far the easiest to implement)" (p. 47). A practical advantage of this parametric approach to the Behrens–Fisher problem is that the Welch–Satterthwaite solution provides a good approximation of the *t*-test significance level under heteroscedasticity (Brunner & Munzel, 2000).

The term "*v*-test", also known as t_v, refers to a *t*-test that uses a different variance estimator for each population instead of only the pooled variance. The table of critical values for the *v*-test is the same as for the *t*-test except that the Welch–Satterthwaite correction for the degrees of freedom is applied. Heiser (2006), who also recommended Welch's approximate *t*-solution (t_{ws}), noted that "if the assumption of normality is valid, then the best method is the *v*-test [using the Welch adjustment for the degrees of freedom]…for all tests on the difference in means, regardless if the variances are equal or unequal" (p. 563).

According to Davenport and Webster (1975), better tests than the Welch–Satterthwaite approximate *t*-test are available to solve the Behrens–Fisher problem. However, because these

other tests are difficult to use and because the Welch–Satterthwaite approximate *t*-test is stable and has acceptable power, they concluded that the latter is preferable. Generally, statisticians use a preliminary test before deciding whether to apply the classical *t*-test or the Welch–Satterthwaite approximate *t*-test to the hypothesis testing of the independent two-sample mean difference. This preliminary test (e.g., Levene's test or the folded form of the *F* statistic, *F*') evaluates the HOV assumption; a *t*-test or approximate *t*-test (Welch–Satterthwaite solution) can then be conducted based on the preliminary test results.

Although the results of the preliminary test and the *t* or approximate *t*-tests are evaluated at the same significance level (α), some researchers advise against this common practice due to the loss of control of the Type I and Type II error rates that occurs when the *t*-test or its approximate version is conducted at the same significance level as the preliminary test used to evaluate the HOV assumption (Heiser, 2006; Sawilowsky, 2002). Sawilowsky (2002), conducted a small study that demonstrate that when the samples are normally distributed, the loss of control of Type I error rate in these cases leads to an inflation that is almost the double of nominal α, when the two tests (i.e., an F test for HOV followed by a *t*-test) are evaluated at a significance level of .05.

The Cochran and Cox *t*-test (Cochran & Cox, 1950) uses the same table of critical values as *t*-test and Welch–Satterthwaite approximate *t*-test. However, the computed *p*-value of the Cochran and Cox *t*-test is based on an adjustment which is different than that of the Welch–Satterthwaite. The Cochran and Cox critical value of *t* is computed as a weighted mean of the two critical values of Sample 1 and Sample 2 (Cochran & Cox, 1950; Lee & Gurland, 1975).

Nonparametric Alternatives

Some nonparametric alternatives have also been proposed for the statistical analysis of the independent two-sample mean difference. One frequently recommended procedure is the Wilcoxon rank-sum test that is commonly used to deal with situations of violations of *t*-test assumptions, including the Behrens–Fisher problem. This test is equivalent to the Mann–Whitney *U* test (Neuhäuser, Lösch, & Jöckel, 2007) and is generally considered to be the nonparametric analogue of the two-independent-samples *t*-test (Howell, 2002; van Belle et al., 2004).

Researchers disagree about this test, however. For example, Stonehouse and Forrester (1998) emphasized that the *U*-test "is not properly a 'non-parametric' analogue of the *t*-test, as it is too often described" (p. 63) because *U* "is not a test of differences between means (and, therefore, not an exact analogue of *t*), but is instead a test of differences between rank orders of samples" (p. 71). In addition, the Wilcoxon–Mann–Whitney (*U*) test seems to be unstable. For example, it is conservative when the sample of larger size comes from the population with the larger variance, and liberal when those factors are reversed (Brunner & Munzel, 2000). Moreover, the *U* test is very sensitive (i.e., non-robust) if the data are both non-normal and heteroscedastic (Stonehouse & Forrester, 1998). This sensitivity is primarily because the *U* test, like the *t*-test, assumes equal variances in the two populations (Kasuya, 2001).

According to Kasuya (2001), the Type I error rate of the *U* test inflates if there is a violation of HOV. This inflation occurs because the Mann–Whitney *U* test assumes a common population for the two samples, (i.e., HOV). Based on the preponderance of evidence against the *U* as an alternative test in situations of violation of HOV assumption, Kasuya concluded that the *U* is not an appropriate test to be used in the presence of heteroscedasticity.

Resampling Approaches

Other approaches that have been attempted or at least suggested as possible solutions to the Behrens–Fisher problem include resampling methods or resampling statistics (Diaconis & Efron, 1983; Efron, 2001; Good, 2005; Howell, 2002). These procedures have been developed relatively recently, in part because they require powerful computing capabilities that were not available 30 or 40 years ago.

Clearly, there are different kinds of resampling methods. I have chosen to focus principally upon two of the most commonly used ones: the permutation test and the nonparametric bootstrap test. These procedures, which are the most popular resampling methods in use today (Alba Fernández et al., 2008), are somewhat related and similar, although there are also outstanding differences between them. One important property they share is their ability to "yield consistent estimators for the distribution of sample means" (Alba Fernández et al., 2008, p. 3731).

The permutation tests and the nonparametric bootstrap methods are nonparametric statistical methods because they rely mostly on empirical analysis to make inferences about the population, based on the observed sample. Although the parametric method also uses the observed sample to make inferences about the population, it relies on parametric theory to obtain the sampling distribution instead of using the empirical sampling distribution, as is done with the resampling methods, to model sampling error (Beasley & Rodgers, 2009). Instead, the sampling distribution in both the permutation tests and the nonparametric bootstrap is created by the researcher's use of different types of computer simulations from the particular set of observed data (i.e., the sample). This use of various types of simulation explains why these, as well as other resampling methods, are known as computer-intensive methods.

The main difference between the permutation tests and the bootstrap lies in how the resampling is conducted. In the permutation tests, the resampling is done using the "without replacement" approach, whereas bootstrap, the samples are obtained using the "with replacement" approach. The main advantage of these two methods is that they do not rely on parametric theory; for this reason, most of the parametric-specific assumptions do not apply to them. They are therefore more flexible, both in terms of applications and interpretation of the results. Nevertheless, after careful study by many researchers, the general conclusion is that this flexibility within resampling methods does not diminish their powerful capacity in comparison to the parametric tests. In fact, at times these resampling methods surpass the parametric tests. In general, resampling methods can be used even if a parametric test application is also appropriate. But they are more appropriate or useful when the more-restrictive assumptions of the parametric tests are violated or cannot be reasonably evaluated, when the sample size is not big enough to obtain meaningful parametric test results, and when no parametric theory or test for a particular statistical problem is available.

When a Behrens–Fisher problem (i.e., violation of the assumption of homogeneity of variance) is present in data, such as that of the sample exercise presented in Chapter 1, it is reasonable to ask whether it is possible to use any of the resampling methods briefly described in this chapter to obtain meaningful results, and whether any of these methods qualify as alternative approaches to the Behrens–Fisher problem. Before attempting to answer these questions, I must briefly describe the permutation tests and the nonparametric bootstrapping approach.

Permutation tests. Permutation tests, sometimes referred as randomization tests, were first described as a method for exact inference by Fisher in 1935 (Ernst, 2004). "The basic idea behind permutation methods is to generate a reference distribution by recalculating a statistic for

many permutations of the data" (Ernst, 2004, p. 676). Today this method is used mostly to test null hypotheses (Beasley & Rodgers, 2009). According to Good (2005), the permutation test's advantage over the parametric *t*-test is that, even for very small samples, the permutation test is exact instead of providing an approximate solution regardless of the normality-assumption status. As a solution for the Behrens–Fisher problem, the main issue with this test is the fact that it contains a disadvantage similar to the parametric two-independent samples *t*-test: both tests rely on the homogeneity of variance assumption (Good, 2005). That is, the naïve application of the permutation test when there is a Behrens–Fisher problem is contrary to the goal of obtaining meaningful results for the testing of the equality of two independent means. It is contrary because the root of the Behrens–Fisher problem is actually the presence of the heterogeneity of variances issue.

As Hayes (2000) emphasized, the assumption of homoscedasticity (i.e., equal population variances or HOV) cannot necessarily be relaxed in the randomization tests for a hypothesis of the equality of two-sample means. Moreover, "there are mathematical arguments that show that a randomization test can be invalid when the population variances are unequal" (Hayes, 2000, p. 654). Hayes (2000) also presented empirical evidence based on Monte Carlo simulations about the invalidity of randomization tests results of the equality of two-sample means when there is a violation of the HOV assumption. He found that when there was a combination of inequalities in both, sample-size and variance in the two groups, the randomization test was either liberal or conservative (Hayes, 2000). His conclusion, with respect to the use of this procedure as a solution to the Behrens–Fisher problem, was that "the randomization test is not necessarily a complete solution to problems produced by the violation of assumptions in the *t* or ANOVA context" (Hayes, 2000, p. 655).

Other non-naïve permutation test approaches are worthy of mention, however. These have been developed by statistical researchers to produce other approximate solutions to the Behrens–Fisher problem. One such solution, proposed by Pesarin (2001), provided what he called an "almost exact solution" to the Behrens–Fisher problem; it is also known as nonparametric combination or NPC (Pesarin & Salmaso, 2010). In addition, Westfall and Young (1993) developed computer algorithms for *p*-value adjustments based on permutation tests under several conditions. These other, more complicated permutation approaches are not generally available in statistical software and will not be used or described further in this dissertation.

The bootstrapping approach. The bootstrapping approach is similar in some ways to the permutation test, except for the sampling replacement characteristic (Efron & Tibshirani, 1993). "In statistics, 'bootstrapping' refers to making inferences about a sampling distribution of statistic by 'resampling' the sample itself with replacement, as if it were a finite population. To the degree that the resampling distribution mimics the original sampling distribution, the inferences are accurate" (Chernick & LaBudde, 2011, p. xi).

It is known that asymptotic tests sometimes overreject or underreject the null hypothesis. However, "by using the bootstrap test instead of an asymptotic one, we can usually, but not always, make more accurate inferences" (MacKinnon, 2002, p. 621). Moreover, according to MacKinnon (2002), "by using bootstrap tests, we may be able to avoid the gross errors of inference that frequently occur when we act as if test statistics actually follow their asymptotic distributions" (p. 623). Both theory and empirical findings also indicate that one of the practical advantages of the bootstrap methods is that they can result in a better control of Type I errors than the non-bootstrap methods can (Keselman, Wilcox, Othman, & Fradette, 2002).

The difference between bootstrap and permutation tests is that the latter are based on the symmetry between the two populations, but the bootstrap only estimates the probability mechanism under the null hypothesis (Efron & Tibshirani, 1993). In addition, the achieved significance level or ASL (i.e., the *p*-value) of the permutation test is exact; by contrast, because of the symmetry, the ASL of the bootstrap is only asymptotically exact, that is, "guaranteed to be accurate as the sample size goes to infinity" (Efron & Tibshirani, 1993, p. 223).

Bootstrapping has become increasingly popular for conducting tests of statistical hypotheses (Davidson & MacKinnon, 2000). One of the common ways to conduct bootstrapping-based hypothesis testing is to compute a *p*-value, because it is the simplest approach to analyze (Davidson & MacKinnon, 2000). Moreover, according to Beasley and Rodgers (2009), the bootstrap is used more often than the permutation test within applied research because the capabilities of the former have surpassed the latter. For example, bootstrap can be applied to hypothesis testing when other resampling methods, such as permutations, are not appropriate. Beasley and Rodgers (2009) also acknowledged a different opinion, namely that of Pesarin (2001), who opined that "for finite sample sizes, inferential interpretations of bootstrap tests are not completely clear" (pp. 126–127) and that "permutation tests are of exact size α … [and] make conditional and unconditional inference interpretation effective and essentially clear" (p. 127). The main advantage of bootstrap over permutation in the two-sample problem is that the latter can be used only to test the null hypothesis of equality of two distributions, F and G (i.e., F = G), whereas the former can be used to test the null hypothesis testing that F = G as well as equal means, with or without the HOV assumption (Efron & Tibshirani, 1993).

Research Questions

In light of literature reviewed herein, a general research question could be: Is either of the two described resampling methods a better approach to the Behrens–Fisher problem than the Welch–Satterthwaite approximate *t*-test or t_{WS}? After studying the evidence in the literature about the uses and assumptions of each of these two resampling methods, as well as the advantages and disadvantages of each, I decided to study the nonparametric bootstrap, in contrast to the Welch–Satterthwaite approximate *t*-test, as an alternative to the Behrens–Fisher problem. Specifically, I decided to conduct a parallel study of three approaches (i.e., two nonparametric bootstrap methods and the Welch–Satterthwaite approximate *t*-test) using computer-simulated data. In addition, I decided to compare and contrast the results of the two nonparametric bootstrap approaches not only with the most commonly found parametrical alternative in today's statistical software packages (i.e., Welch–Satterthwaite solution), but also with other commonly available tests. Those are the classical *t*-test (also known as pooled or Student's) and the Cochran–Cox *t*-test.

The present study was designed to answer two specific questions about a Behrens–Fisher problem in the analysis of two independent normally distributed sample means:

1. Can the nonparametric bootstrap methods as described in Chapter 3, the Welch–Satterthwaite approximate *t*-test, the classical *t*-test, and the Cochran–Cox *t*-test, yield accurate *p*-values for two-tailed hypothesis test, given each of the conditions of this study?

2. Is the hypothesis test based on the proposed nonparametric bootstrap methods, the classical *t*-test, and the Cochran–Cox *t*-test, more powerful[4] than the Welch–Satterthwaite approximate *t*-test, given each of the conditions of this study?

As already stated, the general purpose of this study is to compare and contrast these five different approaches, methods, or tests under a number of different conditions. In Chapter 3, I describe in greater detail the nonparametric bootstrapping methods that I studied. I also describe the conditions (i.e., the combination of different population parameters and sample sizes) that I used for the parallel study.

[4] In this case, the term "powerful" refers to the statistical power of the hypothesis test, which is conventionally defined as 1 - β or 1 - probability of Type II error.

Chapter 3: Methods and Procedures

The design of this study, as described below, is very similar to that of Kohr and Games (1974). An orthogonal design was used to contrast the achieved significance level (*p*-value) and power of five procedures for testing the mean difference of two groups when the HOV could not be assumed. These five statistical procedures, here ordered alphabetically, were the Cochran and Cox *t*-test (hereafter, C&C) assuming unequal variances, a nonparametric bootstrapping method using the Efron and Tibshirani (1993) approach (hereafter, E&T), a nonparametric bootstrapping method using a modified version of the Good (2005) approach, a pooled *t*-test or classical *t*-test, and the approximate *t*-test with a Welch–Satterthwaite (hereafter, W&S) correction.

As described in Chapter 2, the C&C (Cochran & Cox, 1950) uses the same table of critical values as the classical *t*-test but the C&C approximate *t* statistic is computed as a weighted mean of the two critical values of Sample 1 and Sample 2. The E&T (Efron & Tibshirani, 1993) and the modified version of the Good (2005) bootstrap approaches will be described later in this chapter. The pooled *t*-test (also known as classical or Student's *t*-test) is the uniformly most powerful unbiased (UMPU) test for the equality of two independent population means. One of its main features is that it used an estimate of the population variance based of the pooled combination of the two observed variances. Therefore, it is the UMPU test for the equality of two independent population means but only if the assumptions of normality and HOV are not violated.

The Welch–Satterthwaite approximation uses a different estimate for each variance instead of applying the pooled variance to compute the *t*-statistic. As described in Chapter 2 and

according to the literature, a practical advantage of this parametric approach to the Behrens–Fisher problem is that the W&S solution provides a good approximation of the *t*-test significance level under heteroscedasticity. The evaluation of each procedure was conducted in three dimensions (i.e., sample sizes, sample variances and men differences) using SAS software[5]. The same data steps and procedures, as required by this software, were used throughout the present study.

Samples

The samples from a normal population, with mean 0 and variance 1, were generated using different seeds for each of the studied condition-replication-group combination. The seeds were obtained using the SEEDGEN macro as described by Fan, Felsövályi, Sivo, and Keenan (2002). According to these authors, this macro "generates seed values to produce non-overlapping streams of random numbers" (p. 38).

The seeds generated with the SEEDGEN macro were used as the seeds for the STREAMINIT routine of the statistical software, with the RAND (normal) function, to return a different set of random variates from a normal distribution, given a particular combination of condition, replication, and group. For example, the seed of the first simulated sample for Group 1, given its particular conditions (i.e., sample size, mean, and variance), was different than the seed of the first simulated sample for Group 2 given its particular conditions (i.e., sample size, mean, and variance). Therefore, the simulated data of every Group 1 (i.e., Sample 1) on each replication should contain a different set of elements than simulated data of every Group 2 (i.e., Sample 2) on each replication. In the present study, the number of replications for each condition was 2,000.

[5] SAS software (Versions 9.3) was used for all analyses described in of this section.

Dimensions (Sample Sizes, Sample Variances, and Mean Differences)

Type I error rate. All samples of size (n) were drawn randomly, with replacement, from the generated population. The sample size of Group 1 had three levels (i.e., n_1 = 10, 25, and 40). The proportionality of the two sample sizes was represented by five levels of sample-size ratios (i.e., n_2/n_1 = 1, 1.5, 3, and 5), rounding the sample size of group 2 (n_2) to the nearest integer:

The variance of group 2 (Var_2) was fixed to 1 throughout the present study, while the condition of heterogeneity of variance was represented by five levels of variance ratios (Var_1/Var_2 = 1/16, 1/4, 1/2, 1, 2, 4, and 16). Therefore, only the variance of Group 1 (Var_1) was manipulated to obtain the desired variances ratio. This manipulation consisted of multiplying every randomly drawn value of the first group by the desired variance ratio.

The combination of the sample size conditions and the variance ratio conditions produced the three dimensions of the study of Type I error5 rate. These dimensions were crossed (3 x 4 x 7). Therefore, a total of 84 conditions were investigated for the Type I error rate portion of this study.

Power analysis. Four equally spaced points (i.e., mean differences) on the power curve were established for each condition. Four sample means (M) of Group 1 (M_1 = 0.5, 1.0, 1.5, and 2.0) were evaluated. Therefore, to obtain the desired mean of group 1 (M_1), the corresponding constant, as shown here, was added to every randomly drawn value of the first group. The mean of group two (M_2) was fixed to 0 throughout the present study; therefore, the true mean differences, $M_1–M_2$, were always positive, corresponding to the four values of M_1[6].

[6] To test if a negative difference (i.e., $M_1–M_2 < 0$), would make any difference in the results, some of the extreme combinations of variance and sample size ratios were evaluated by inverting the sign of M_1 from positive to negative. Those results are presented in Appendix D. As Tables D1 to D6 show, when the sign of M_1 was inverted from positive to negative, the results were very close to those when the difference was positive (i.e., $M_1–M_2 > 0$), except for some occasionally small sampling and/or rounding errors. Therefore, it seems to be safe to conclude that there is no difference in the results.

The combination of the sample size conditions, variance ratio conditions, and mean differences were crossed (3 x 4 x 7 x 4). Therefore, a total of 336 conditions were investigated for the power analysis.

The Two Criteria of the Study: Type I Error Rate (*p*-value) and Power of Each Method

The present study was conducted based on two major criteria: Type I error rate (i.e., achieved significance level or *p*-value) and power analysis (i.e., empirical estimation of power). Note that the term "standard" refers to the significance level of the null hypothesis of the methods at .05 and .01. In other words, the standards were the nominal significance levels (i.e., $\alpha = .05$ and $\alpha = .01$) of the hypothesis tests.

Parametric methods (i.e., methods based on *t*-test). The following represents the general algorithm used to compute the Type I error rate and the power estimate of the three parametric methods studied.

1. Create six counters, one for each method-significant level combination (i.e., pooled *t*-test_.01, pooled *t*-test_.05, C&C *t*-test_.01, C&C *t*-test_.05, W&S *t*-test_.01, and W&S *t*-test _.05) that will contain the number of times each result for each method-standard combination was significant.
2. Set all counters to 0.
3. Simulate a different set of data for every replication, given the appropriate sample size ratio, sample variance ratio, and mean difference.
4. Use the TTEST procedure of the statistical software to obtain the *p*-value of the *t*-test for the three parametric-based methods studied (i.e., pooled or classical *t*-test, W&S, and C&C).

5. Evaluate if each of the *p*-values obtained in Step 4 is significant to .01. If so, increase the corresponding counter by 1. For example, if the *p*-value of the pooled *t*-test was less or equal to .01, the counter-pooled *t*-test_.01 is increased by 1.

6. Evaluate if each of the *p*-values obtained in Step 4 is significant to .05. If so, increase the corresponding counter by 1. For example, if the *p*-value of the classical *t*-test was less or equal to .05, the counter-classical *t*-test_.05 is increased by 1.

7. Repeat steps 3 to 6 for the number of total replications (i.e., 2,000).

8. Divide each of the 10 counters by 2,000 (i.e., the number of replications). These values represent the detection rates. These detection rates are the Type I error rates for the set of simulated conditions when H_0 was true ($\mu_1 = \mu_2 = 0$). When H_0 was false ($\mu_1 \neq \mu_2$), these detection rates are the empirical estimates of power for the set of simulated conditions.

9. Repeat steps 1–8 for each set of simulated conditions.[7]

Bootstrap-based methods. The nonparametric bootstrapping approach for the hypothesis testing of the two-sample mean (i.e., H_0: $\mu_1 = \mu_2 = 0$) could be conducted in several ways (Chernick & LaBudde, 2011, Efron & Tibshirani, 1993; Good, 2005). The ASL or *p*-value of the two nonparametric bootstrap approaches considered in the present study corresponds to the estimated achieved significance level or empirical *p*-value computed using Good's (2005) and Efron and Tibshirani's (1993) approaches.

Modified Good bootstrap. Good's (2005) testing approach is conducted using a bootstrap-based confidence interval. However, he suggested that the samples to construct those confidence intervals should be obtained by drawing "two separate bootstrap samples each time, one from each of the original two samples" (p. 46). The test statistic in the case of the modified

[7] There is a total of 420 (3 x 4 x 7 x 5) sets of simulated conditions in this study.

version of the Good (2005) approach is the presence of the null value (i.e., 0) within the nonparametric bootstrap interval of the mean difference, obtained with the simple percentile method. Therefore, the ASL or empirical *p*-value consists of the number of times, out of the total number of bootstrap replications that the null value lies outside the bootstrap confidence interval. More specifically, the two-tailed Good-based Type I error rate was obtained using the Efron and Tibshirani (1993) approach of computing the percentile bootstrap confidence interval and counting how many times the value 0 was not part of the intervals (i.e., the significant results).

The following represents the general algorithm used to compute the Type I error rate and the power estimate of a modified version of the Good bootstrap method based on bootstrap confidence intervals.

1. Create four counters, two for each condition-significant level combination (i.e., Good_.05$_a$, Good_.05, Good_.01$_a$, and Good_.01. The counters with the subscript (i.e., a), was be used only to temporarily accumulate the number of times that each individual bootstrap replication, given its standard (i.e., .05 or .01), is significant. The other two counters (i.e., those without the subscript) will permanently accumulate the number of times the ASL or empirical *p*-value of the bootstrap sample, given each standard, was significant. These latter two counters are analogous to those used with the parametric methods to obtain the detection rates.
2. Set all counters to 0.
3. Simulate a different set of data for every replication, given the appropriate sample size ratios, sample variance ratios, and mean differences.

4. Obtain a bootstrap data set (x*, y*) from the data simulated in Step 3, where x* corresponds to a sample with replacement from Group 1 and y* corresponds to a sample with replacement from Group 2.[8]

5. Compute the difference between the means of x* and y*.

6. Compute the percentile bootstrap confidence interval from the differences computed in step 5, at a significance level of .05.

7. Evaluate the significance of the percentile bootstrap confidence interval at a significance level of .05 to see if the value 0 was not part of the interval (i.e., the significant result). If it was not, increase the Good_.05_a counter by 1.

8. Compute the percentile bootstrap confidence interval from the differences computed in step 5, at a significance level of .01.

9. Evaluate the significance of the percentile bootstrap confidence interval at a significance level of .01 to see if the value 0 was not part of the interval (i.e., the significant result). If it was not, increase the Good_.01_a counter by 1.

10. Repeat steps 4 through 8 b times. The number represented by b corresponds to the number of bootstrap replications (e.g., 9,999).

11. The ASL or empirical *p*-value of the bootstrap sample at each significance level consists of the number of times, out of b, that the percentile bootstrap obtained in Step 6 for a significance level of .01 and 8 for a significance level of .05, were significant (i.e., the percentile

[8] Note that each sample, x* and y*, is drawn separately from its respective group.

bootstrap confidence interval did not contain the value 0). In other words, divide counters Good_.05$_a$ and Good_.01$_a$ by 9,999[9].

12. If the ASL or empirical *p*-value of the bootstrap sample at a significance level of .05 is less than .05, increase Good_.05 by 1.

13. If the ASL or empirical *p*-value of the bootstrap sample at a significance level of .01 is less than .01, increase Good_.01 by 1.

14. Reset Good_.05$_a$ and Good_.01$_a$ counters to 0.

15. Repeat steps 3 to 13 according to the number of total replications (i.e., 2,000).

16. Divide the counters Good_.05 and Good_.01 by 2,000 (i.e., the number of replications). These values represent the detection rates. These detection rates are the Type I error rates for the set of simulated conditions when H_0 was true ($\mu_1 = \mu_2 = 0$). On the other hand, when H_0 was false ($\mu_1 \neq \mu_2$), these detection rates are the empirical estimates of power for the set of simulated conditions.

17. Repeat steps 1–15 for each set of simulated conditions.

Efron and Tibshirani bootstrap. According to Efron and Tibshirani (1993) "more accurate testing can be obtained through the use of a studentized statistic" (p. 221). These authors also described an algorithm to obtain the bootstrap-based estimated ASL when the variances of the two independent groups are not equal, a Behrens-Fisher problem (Efron & Tibshirani, 1993).

However, the test statistic of the algorithm proposed by Efron and Tibshirani (1993) for "computation of the bootstrap test statistic for testing equality of means" (p. 224) is based on the

[9] In this study the "99 rule" (Beasley & Rodgers, 2009; Boos, 2003) was used to compute the empirical *p*-value. Therefore, instead of using Good_.01a / 9,999 and Good_.05a / 9,999, the empirical *p*-values were computed using the following correction: (Good_.01a + 1) / (9,999 + 1) and (Good_.05a + 1) / (9,999 + 1), respectively.

approximate *t*-test, which corresponds to the *t*-statistic when the homogeneity of variances (HOV) is not assumed. The Efron and Tibshirani approach also involves centering of the scores around a common mean as well as drawing two separate bootstrap samples, as described below.

The following represents the general algorithm used to compute the Type I error rate and the power estimate of the Efron and Tibshirani (1993) bootstrap method based on the bootstrap of *t*-test.

1. Create two counters, for each condition-significant level (i.e., ET_.05 and ET_.01). Additionally, create another counter for the significance of the bootstrap *t*-test (i.e., ET_sig).

2. Set the three counters to 0.

3. Simulate a different set of data for every replication, given the appropriate sample size ratios, sample variance ratios, and mean differences.

4. Compute the approximate two-sample independent *t*-test (i.e., a *t*-test without assuming equal variance) from the original set of data obtained from Step 3.

5. Translate or center the mean of each group independently around the mean of the combined sample.

a. For each group, subtract from each observation its respective group mean.

b. Add to each observation the mean of the combined sample. Thus, both groups will have the same mean, which corresponds to the mean of the combined sample.

6. Obtain a bootstrap data set (x*, y*) from the data translated in Step 5, where x* corresponds to a sample with replacement from Group 1 and y* corresponds to a sample with replacement from Group 2.[10]

[10] Note that each sample, x* and y*, is drawn separately from its respective group.

7. Compute the unequal variances two-sample *t*-test using the bootstrap data set obtained in Step 5.

8. Evaluate whether the *t*-test value obtained in Step 7 is greater than the *t*-test of the original set of data obtained from Step 4. If it is greater, increase counter ET_sig by 1.

9. Repeat steps 6–8 b times. The number b corresponds to the number of bootstrap replications (e.g., 9,999).

10. The ASL or empirical *p*-value consists of the number of times, out of b, that the *t*-test values obtained in Step 7 were greater than the value of the observed t_{ws} obtained from Step 4. In other words, divide counter ET_sig by 9,999.[11]

11. Evaluate whether ET_sig is less than or equal to .05. If it is, increase ET_.05 by 1.

12. Evaluate whether ET_sig is less than or equal to .01. If it is, increase ET_.01 by 1.

13. Reset ET_sig to 0.

14. Repeat steps 3–13 for the number of total replications (i.e., 2,000).

15. Divide the counters ET_.05 and ET_.01 by 2,000 (i.e., the number of replications). These values represent the detection rates. These detection rates are the Type I error rates for the set of simulated conditions when H_0 was true ($\mu_1 = \mu_2 = 0$). When H_0 was false ($\mu_1 \neq \mu_2$), these detection rates are the empirical estimates of power for the set of simulated conditions.

16. Repeat steps 1–15 for each set of simulated conditions.

Summary

Five of the methods suggested in the literature as approaches to the Behrens–Fisher problem were studied, using simulated data under different set of dimensions (i.e., simulated

[11] In this study, the "99 rule" (Beasley & Rodgers, 2009; Boos, 2003) was used to compute the empirical *p*-value. Therefore, instead of using ET_Sig / 9,999, the empirical *p*-values were computed using the following correction: (ET_Sig + 1) / (9,999 + 1).

conditions for sample sizes, sample variances, and mean differences). Three of those methods are parametric-based alternatives (i.e., C&C *t*-test assuming unequal variances, pooled *t*-test or classical *t*-test, and the approximate *t*-test with a W&S correction); the other two are bootstrap-based (i.e., a nonparametric bootstrapping method using the E&T approach and a nonparametric bootstrapping method using a modified version of the Good approach). In Chapter 4, I present the results of the simulations of the five methods described above. In Chapter 5, I discuss the results presented in Chapter 4.

Chapter 4: Results

The results of this study are based on computer-simulated data of two groups from normally distributed populations. The data were simulated using different combinations of conditions or sets of parameters. A Monte Carlo experiment of 2,000 replications was conducted that applied the different sets of parameters for a Type I error rate study and for a power analysis. For the study of the Type I error rates, the conditions were the sample size of Group 1 (n_1), the proportionality of the two sample sizes or the sample-size ratio (n_2/n_1), and the variance ratios (var_1/var_2) of the two groups. For the power analysis, four equally spaced points (i.e., mean differences) on the power curve were established for each combination of conditions or sets of parameters of the simulated data.

Five methods were studied. The methods, in alphabetical order, were the Cochran and Cox *t*-test (C&C); the Efron and Tibshirani non-parametric bootstrap (E&T); the Good bootstrap; the pooled *t*-test, also known as Gosset/Student or classical *t*-test; and the Satterthwaite *t*-test[12]. Each of the 2,000 bootstrap replications was based on 9,999 bootstrap samples. All of the analyses (i.e., hypothesis testing of equality of the two means with each method) were evaluated at two standards (i.e., significance levels), .05 and .01. The presentation of the results is divided into two main sections: Type I error rates and power analysis.

[12]As described in Chapter 2, the Satterthwaite and the Welch solutions are equivalent. That is why this approximate *t*-test solution to the Behrens–Fisher problem is commonly known as the Welch–Satterthwaite solution (WS). In this chapter, this solution will be referred to as the Satterthwaite approximate *t*-test solution because the statistical software package used in this study, uses the Satterthwaite solution by default.

Type I Error Rates

As discussed in Chapter 3, Type I error rates are proportions of statistically significant results out of *k* replications when the null hypothesis (H_0) is true (i.e., $\mu_1 = \mu_2$). The number of replications in the present study was 2,000. Once those proportions were computed, the next step was to evaluate if they are statistically significantly different from their respective standards of .05 and .01. Given that the number of individual experiments in this study was 2,000, a normal approximation to the binomial test could be used to evaluate the significance of each proportion. However, before assessing the statistical significance of each proportion, the significance level for the normal approximation to the binomial test had to be adjusted.

It is well known that "The more tests we perform on a set of data, the more likely we are to reject the null hypothesis when it is true. […] This problem is called the inflation of the alpha level" (Abdi, 2007, p. 103). An inflation of the alpha (α) level risks the false conclusion that there are significant statistical results when there are really none (Abdi, 2007). In the present study, an inflation of the alpha level could lead us to think that a particular test result is statistically significant when it really is not. One way to deal with this situation, which is also known as a multiplicity problem or issue, involves making the alpha level lower or more stringent (Abdi, 2007). This approach, in which the alpha level of a hypothesis test is adjusted to correct the multiplicity effect, is sometimes referred to as a Bonferroni-type correction. In the present study, instead of evaluating the results of the binomial test for the proportions of significant results at the alpha of the original hypothesis tests, a more stringent alpha of .001 was used.

Table 1 contains the lower and upper bounds of the confidence interval of the normal approximation of the binomial proportions of the significant results, based on 2,000 replications

of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true and the standards (i.e., the significance levels of the hypothesis tests of equality of the means) were .01 and .05. If an observed proportion is not within its respective interval for the standard, then it is a statistically significant result.

Table 1

Lower and upper bounds of confidence interval of the binomial proportion of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards (i.e., the significance levels of the hypothesis tests of equality of the means) were .01 and .05; at a significance level of .001

Standard	Lower bound	Upper bound
.05	0.0340	0.0660
.01	0.0027	0.0173

As Table 1 shows, when the standard (i.e., nominal α) of the hypothesis test was .05, any proportion of significant results (i.e., Type I error rate) smaller than 0.0340 or greater than 0.0660 is considered statistically significant. Similarly, when the standard or nominal α was .01, any proportion of Type I error rate smaller than 0.0027 or greater than 0.0173 is also considered statistically significant. Note that although the statistical significance of each proportion in the present study was formally evaluated with the *p*-value results of the binomial test, using the FREQ procedure of the statistical software, the conclusions in terms of the significance of the proportions are the same. In other words, the assessment of the significance of the binomial proportions was expected produce the same conclusion if either the confidence interval of the normal approximation of binomial proportion approach (Table 1) or the *p*-value results of the binomial test were used.

The following presentation of the Type I error rate results is divided into subsections based on the sample size of the first group (n_1) and the sample-size ratios (n_2/n_1). Within each subsection, the level of Type I error rate control of each particular method is classified into five

categories based on the number of significant proportions at a significance level (α) of .001, given a simulated sample-size ratio. The five categories are: *completely controlled*, when all of the observed results are statistically non-significant; *relatively well controlled*, when only one or two observed results are statistically significant; *moderately controlled*, when three observed results are statistically significant; *poorly controlled*, when four or five observed results are statistically significant; and *not controlled*, when six or seven observed results are statistically significant. As previously specified, the discussion of the Type I error rates will focus on the results based on the more stringent adjusted significance level (α). In other words, only those results that were significant at α level for the binomial test of the proportion of .001 (Table 1) will be discussed.

Sample size group 1 (n_1) = 10 and equal sample sizes ($n_1 = n_2 = 10$). Table 2 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes were equal (i.e., $n_1 = n_2 = 10$) for all the variance ratios, and the standards (i.e., nominal α or the significance levels of the hypothesis tests of equality of the means) were .05 or .01. The results[13] (i.e., empirical α or Type I error rate) of all methods are presented.

As shown in Table 2, when the standard was .05, the pooled *t*-test and the Satterthwaite *t*-test showed complete Type I error rate control (i.e., the nominal α was statistically equal to the empirical α) in all of the variance ratios. Two methods that showed relatively good control of Type I error rates were the C&C and the E&T bootstrap. The C&C method failed to control (i.e., the nominal α was statistically significant different than the empirical α) in only two cases: when the variance ratios were 2 and 4. The E&T bootstrap method failed to control on only one

[13] In this study, the Type I error rates, rounded to four decimal places, are based on 2,000 replications of the simulated experiment.

occasion, when the sample variance ratio was 4. The Good[14] bootstrap method failed to control for Type I error rates on all occasions.

Table 2

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 10; sample-size ratio (n_2/n_1) was 1.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0475	0.0430	0.0980*	0.0635	0.0545
1/4	0.0415	0.0425	0.0800*	0.0525	0.0490
1/2	0.0460	0.0550	0.0920*	0.0595	0.0580
1	0.0365	0.0430	0.0725*	0.0465	0.0455
2	0.0320*	0.0360	0.0715*	0.0410	0.0380
4	0.0295*	0.0335*	0.0720*	0.0435	0.0395
16	0.0425	0.0370	0.0870*	0.0540	0.0465
			Standard .01		
1/16	0.0085	0.0080	0.0370*	0.0185*	0.0125
1/4	0.0065	0.0080	0.0320*	0.0110	0.0105
1/2	0.0040	0.0065	0.0330*	0.0125	0.0105
1	0.0040	0.0080	0.0295*	0.0110	0.0110
2	0.0040	0.0075	0.0245*	0.0095	0.0090
4	0.0060	0.0055	0.0215*	0.0105	0.0090
16	0.0070	0.0060	0.0330*	0.0135	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

When the standard was .01, the C&C, E&T bootstrap, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 2). The pooled *t*-test method showed relatively good control of Type I error rates. It failed to control for Type I

[14] From this point forward, the terms *Good* and *modified version of the Good* are used interchangeably because for the purpose of this study, both terms refer to the same bootstrap method or approach.

error rates on only one occasion: when the variance ratio was 1/16. The Good bootstrap method failed to control for Type I error on all occasions.

Sample size group 1 (n_1) = 10 and sample-size ratio (n_2/n_1) = 1.5. Table 3 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 1.5. Therefore, the sample size of the first group (i.e., n_1) was 10 and the sample size of the second group (i.e., n_2) was 15. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

As shown in Table 3, when the standard was .05, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The pooled *t*-test method showed moderately controlled Type I error rates. It failed to control on three occasions: when variance ratios were 1/4, 4, and 16. The Good bootstrap method failed to control for Type I error rates on all occasions.

Similarly to the cases when the standard was .05, when the standard was .01 the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 3). The pooled *t*-test method showed relatively good control for Type I error rates. It failed to control on two occasions: when variance ratios were 4 and 16. The Good bootstrap method failed to control for Type I error on all occasions.

Table 3

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 10; sample-size ratio (n_2/n_1) was 1.5; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0590	0.0600	0.0925*	0.0355	0.0610
1/4	0.0420	0.0505	0.0745*	0.0320*	0.0530
1/2	0.0450	0.0545	0.0820*	0.0460	0.0565
1	0.0380	0.0435	0.0720*	0.0465	0.0440
2	0.0395	0.0440	0.0765*	0.0640	0.0470
4	0.0365	0.0355	0.0770*	0.0760*	0.0410
16	0.0505	0.0425	0.0960*	0.1140*	0.0525
			Standard .01		
1/16	0.0095	0.0105	0.0350*	0.0045	0.0135
1/4	0.0060	0.0090	0.0225*	0.0050	0.0110
1/2	0.0045	0.0080	0.0275*	0.0065	0.0110
1	0.0075	0.0085	0.0230*	0.0110	0.0100
2	0.0050	0.0060	0.0275*	0.0170	0.0075
4	0.0035	0.0045	0.0290*	0.0205*	0.0075
16	0.0080	0.0075	0.0405*	0.0395*	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

Sample size group 1 (n_1) = 10 and sample-size ratio (n_2/n_1) = 3.0. Table 4 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 3.0. Therefore, the sample size of the first group (i.e., n_1) was 10 and the sample size of the second group (i.e., n_2) was 30. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

Table 4

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 10; sample-size ratio (n_2/n_1) was 3.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0400	0.0430	0.0590	0.0015*	0.0445
1/4	0.0405	0.0525	0.0670*	0.0070*	0.0520
1/2	0.0365	0.0445	0.0640	0.0155*	0.0450
1	0.0375	0.0445	0.0780*	0.0445	0.0465
2	0.0420	0.0425	0.0820*	0.0990*	0.0465
4	0.0465	0.0440	0.0880*	0.1535*	0.0490
16	0.0415	0.0295*	0.0845*	0.2230*	0.0415
			Standard .01		
1/16	0.0050	0.0065	0.0150	<0.0005*	0.0075
1/4	0.0045	0.0080	0.0140	0.0010*	0.0090
1/2	0.0045	0.0055	0.0160	0.0015*	0.0065
1	0.0065	0.0075	0.0200*	0.0100	0.0080
2	0.0090	0.0115	0.0285*	0.0275*	0.0140
4	0.0075	0.0070	0.0365*	0.0620*	0.0100
16	0.0075	0.0050	0.0350*	0.1055*	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 4, when the standard was .05, the C&C and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The E&T bootstrap method showed relatively well controlled Type I error rates. It failed to control only when the variance ratio was 16. The Good bootstrap method failed to control for Type I error rates on five occasions. The pooled *t*-test failed to control in all occasions, except when the variances were equal (i.e., variance ratio was 1.0). Therefore, the Good bootstrap method showed poor control whereas the pooled *t*-test showed no control over Type I error rates.

When the standard was.01, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 4). The Good bootstrap method failed to control for Type I error rates on four occasions, when the variance ratios were 1 or greater. Therefore, its control over Type I error rates was poor. The Type I error rates were not controlled by the pooled *t*-test method, similarly to the cases when the standard was .01. This method failed to control on all occasions, except when the variances were equal (i.e., variance ratio was 1.0).

Sample size group 1 (n_1) = 10 and sample-size ratio (n_2/n_1) = 5.0. Table 5 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 5.0. Therefore, the sample size of the first group (i.e., n_1) was 10 and the sample size of the second group (i.e., n_2) was 50. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

As shown in Table 5, when the standard was .05, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The Good bootstrap method showed poor control over the Type I error rate. It failed to control on almost all occasions, except when variance ratios were 1/4 or 1/16. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1.0). Therefore, it does not control for Type I error rates.

Similarly to the cases when the standard was .05, when the standard was .01 the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 5). The Type I error rate was not controlled by the Good bootstrap and the pooled *t*-test methods. These two methods failed to control in all occasions, except for one. In the case of the Good bootstrap method, the Type I error was

controlled only when the variance ratio was 1/16. In the case of the pooled *t*-test method, the Type I error was controlled only when the variances were equal (i.e., variance ratio was 1).

Table 5

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 10; sample-size ratio (n_2/n_1) was 5.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0360	0.0430	0.0560	0.0005*	0.0435
1/4	0.0345	0.0435	0.0610	0.0040*	0.0450
1/2	0.0435	0.0505	0.0760*	0.0155*	0.0530
1	0.0480	0.0505	0.0885*	0.0565	0.0530
2	0.0410	0.0385	0.0840*	0.1090*	0.0475
4	0.0475	0.0435	0.0980*	0.2050*	0.0490
16	0.0585	0.0515	0.1190*	0.3320*	0.0590
			Standard .01		
1/16	0.0060	0.0085	0.0145	<0.0005*	0.0090
1/4	0.0075	0.0130	0.0210*	0.0005*	0.0125
1/2	0.0105	0.0140	0.0265*	0.0010*	0.0155
1	0.0100	0.0105	0.0290*	0.0125	0.0120
2	0.0085	0.0080	0.0280*	0.0380*	0.0095
4	0.0085	0.0080	0.0385*	0.1000*	0.0105
16	0.0140	0.0090	0.0475*	0.2140*	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

Overall summary of Type I error rates for $n_1 = 10$ and standard = .05. The Satterthwaite *t*-test showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. Two methods that showed relatively good control of Type I error rates were the C&C and the E&T bootstrap. The C&C method failed to control in only two cases: when the samples were equal (i.e., sample-size ratios were

1.0). Similarly, the E&T bootstrap method failed to control on two occasions, but with different sample-size ratios: one when the sample-size ratio was 1.0, and the other when the sample-size ratio was 3.0.

The pooled *t*-test method failed to control for Type I error rates except when the sample sizes were equal; then, it showed complete Type I error rate control. It showed a moderate control when the sample-size ratio was 1.5. Additionally, as expected, the pooled *t*-test method showed Type I error rate control when the variances were equal (i.e., the variance ratio was 1), regardless of the sample-size ratios. The Good bootstrap method failed to control for Type I error rates most of the time.

Overall summary of Type I error rates for n_1 = 10 and standard = .01. The C&C, E&T bootstrap, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. The Good bootstrap method failed to control for Type I error rates most of the time. The pooled *t*-test method also failed to control most of the time, especially when the sample-size ratios were greater than 1.5, but showed relatively good control when the sample-size ratios were 1.0 (i.e., the sample sizes were equal) or 1.5. As expected, the pooled *t*-test method also showed adequate Type I error rate control when the variances were equal (i.e., the variance ratio was 1), regardless of the sample-size ratios.

Sample size group 1 (n_1) = 25 with equal sample sizes ($n_1 = n_2 = 25$). Table 6 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes were equal (i.e., $n_1 = n_2 = 25$) for all the variance ratios, and the standards (i.e., the significance levels of the hypothesis tests of equality of the means) were .05 or .01. The results (i.e., Type I error rates) of all methods are presented.

Table 6

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 25; sample-size ratio (n_2/n_1) was 1.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0455	0.0455	0.0670*	0.0500	0.0480
1/4	0.0380	0.0400	0.0495	0.0415	0.0405
1/2	0.0380	0.0415	0.0505	0.0425	0.0425
1	0.0460	0.0515	0.0635	0.0515	0.0515
2	0.0455	0.0500	0.0620	0.0510	0.0505
4	0.0545	0.0565	0.0735*	0.0620	0.0585
16	0.0420	0.0415	0.0595	0.0465	0.0430
			Standard .01		
1/16	0.0055	0.0045	0.0170	0.0085	0.0060
1/4	0.0080	0.0090	0.0155	0.0100	0.0085
1/2	0.0065	0.0075	0.0100	0.0080	0.0080
1	0.0070	0.0095	0.0120	0.0085	0.0085
2	0.0100	0.0115	0.0170	0.0125	0.0120
4	0.0100	0.0115	0.0220*	0.0125	0.0120
16	0.0060	0.0040	0.0155	0.0085	0.0065

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 6, all methods, except for the Good bootstrap, showed complete Type I error rate control in all of the variance ratios, at both standards (i.e., .05 and .01). The Good bootstrap method showed relatively good control over the Type I error rate at both standards. It failed to control only when variance ratios were 1/16 and 4, at the standard of .05, and when the variance ratio was 4 at the standard of .01.

Sample size group 1 (n_1) = 25 and sample-size ratio (n_2/n_1) = 1.5. Table 7 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the

sample sizes ratio (i.e., n_2/n_1) was 1.5. Therefore, the sample size of the first group (i.e., n_1) was 25 and the sample size of the second group (i.e., n_2) was 38. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

Table 7

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 25; sample-size ratio (n_2/n_1) was 1.5; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0475	0.0490	0.0620	0.0195*	0.0500
1/4	0.0410	0.0460	0.0550	0.0250*	0.0465
1/2	0.0465	0.0540	0.0635	0.0385	0.0535
1	0.0495	0.0540	0.0640	0.0550	0.0530
2	0.0535	0.0560	0.0680*	0.0760*	0.0565
4	0.0530	0.0530	0.0675*	0.0855*	0.0550
16	0.0535	0.0530	0.0760*	0.1190*	0.0535
			Standard .01		
1/16	0.0085	0.0090	0.0155	0.0005*	0.0095
1/4	0.0075	0.0080	0.0120	0.0040	0.0090
1/2	0.0085	0.0105	0.0150	0.0065	0.0105
1	0.0110	0.0125	0.0195*	0.0150	0.0140
2	0.0110	0.0110	0.0175*	0.0195*	0.0110
4	0.0095	0.0105	0.0190*	0.0245*	0.0110
16	0.0120	0.0120	0.0230*	0.0400*	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 7, when the standard was .05, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The Type I error rates were moderately controlled by the Good bootstrap method, which

failed to control only when variance ratios were 2 or greater. The pooled *t*-test method failed to control for Type I error rates on five occasions. Therefore it showed poor control for Type I error rates. However, as expected, it controlled for Type I error rates when the variances were equal (i.e., the variance ratio was 1).

Similarly to the cases when the standard was .05, when the standard was .01, the C&C, the E&T, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 7). The Type I error rate was poorly controlled by the Good bootstrap and the pooled *t*-test methods, which failed to control on four occasions. In the case of the Good bootstrap method, the control over the Type I error failed when the variance ratios were 1 or greater whereas the pooled *t*-test method failed to control when the variances ratios were 1/16, 2, 4, and 16.

Sample size group 1 (n_1) = 25 and sample-size ratio (n_2/n_1) = 3.0. Table 8 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 3.0. Therefore, the sample size of the first group (i.e., n_1) was 25 and the sample size of the second group (i.e., n_2) was 75. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

As shown in Table 8, when the standard was .05, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. In the case of the Good bootstrap method, the Type I error rates were relatively well controlled. It failed to control only when the variance ratio was16. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1).

Table 8

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 25; sample-size ratio (n_2/n_1) was 3.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0485	0.0490	0.0545	0.0025*	0.0485
1/4	0.0475	0.0525	0.0570	0.0085*	0.0505
1/2	0.0500	0.0530	0.0620	0.0235*	0.0540
1	0.0505	0.0515	0.0600	0.0520	0.0525
2	0.0490	0.0485	0.0605	0.0925*	0.0505
4	0.0430	0.0410	0.0590	0.1510*	0.0430
16	0.0515	0.0500	0.0670*	0.2210*	0.0520
			Standard .01		
1/16	0.0095	0.0110	0.0120	<0.0005*	0.0100
1/4	0.0070	0.0095	0.0130	<0.0005*	0.0100
1/2	0.0100	0.0100	0.0130	0.0040	0.0100
1	0.0085	0.0100	0.0180*	0.0105	0.0120
2	0.0070	0.0075	0.0165	0.0300*	0.0080
4	0.0040	0.0035	0.0095	0.0500*	0.0045
16	0.0105	0.0085	0.0225*	0.1080*	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

Similarly to the cases when the standard was .05, when the standard was .01, the C&C, the E&T, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 8). In the case of the Good bootstrap method, the Type I error rates were relatively well controlled. It failed to control only when the variance ratios were 1 and 16. The pooled *t*-test method failed to control for Type I error rates on most occasions, except when the variance ratio was 1/2 and when the variance ratio was 1 (i.e., variances were equal). Therefore, its control for Type I error rates was poor.

Sample size group 1 (n_1) = 25 and sample-size ratio (n_2/n_1) = 5.0. Table 9 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 5.0. Therefore, the sample size of the first group (i.e., n_1) was 25 and the sample size of the second group (i.e., n_2) was 125. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

Table 9

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 25; sample-size ratio (n_2/n_1) was 5.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0525	0.0540	0.0570	<0.0005*	0.0550
1/4	0.0460	0.0510	0.0575	0.0020*	0.0505
1/2	0.0500	0.0515	0.0650	0.0120*	0.0540
1	0.0450	0.0485	0.0600	0.0470	0.0490
2	0.0425	0.0420	0.0605	0.1145*	0.0440
4	0.0480	0.0480	0.0630	0.1955*	0.0495
16	0.0450	0.0435	0.0650	0.3315*	0.0455
			Standard .01		
1/16	0.0095	0.0110	0.0125	<0.0005*	0.0110
1/4	0.0075	0.0080	0.0125	<0.0005*	0.0080
1/2	0.0075	0.0085	0.0170	0.0015*	0.0090
1	0.0065	0.0070	0.0130	0.0085	0.0075
2	0.0075	0.0070	0.0155	0.0375*	0.0075
4	0.0095	0.0090	0.0175*	0.0875*	0.0095
16	0.0085	0.0080	0.0150	0.2000*	0.0085

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 9, when the standard was .05, the C&C, the E&T bootstrap, the Good bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1).

When the standard was .01, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 9). In the case of the Good bootstrap method, the Type I error rates were relatively well controlled. It failed to control only when the variance ratio was 4. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variance ratio was 1 (i.e., variances were equal).

Overall summary of Type I error rates, n_1 = 25 and standard = .05. The C&C, E&T bootstrap, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. The Good bootstrap method showed complete control for Type I error rates when the sample-size ratios were 5.0, relatively good control when the sample-size ratios were 1.0 and 3.0, and moderate control when the sample-size ratios were 1.5. The pooled *t*-test method failed to control for Type I error rates except when the sample sizes were equal or when the sample-size ratios were 1.5. It showed complete Type I error rate control when the sample sizes were equal showed poor control at sample-size ratios of 1.5. As expected, the pooled *t*-test method also showed Type I error rate control when the variances were equal (i.e., the variance ratio was 1.0), regardless of the sample-size ratios.

Overall summary of Type I error rates, n_1 = 25 and standard = .01. The C&C, E&T bootstrap, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. The Good bootstrap

method showed relatively good control for Type I error rates except when the sample-size ratio was 1.5, when it showed poor control. The pooled *t*-test method showed complete control of Type I error rates when the samples were equal (i.e., sample-size ratios of 1.0) but failed to control in most instances as the sample-size ratios increased. As expected, the pooled *t*-test method showed adequate Type I error rate control when the variances were equal (i.e., the variance ratio was 1.0), regardless of the sample-size ratios.

Sample size group 1 (n_1) = 40 and equal sample sizes ($n_1 = n_2 = 40$). Table 10 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes were equal (i.e., $n_1 = n_2 = 40$), for all the variance ratios, and the standards (i.e., the significance levels of the hypothesis tests of equality of the means) were .05 or .01. As shown in Table 10, all methods showed complete Type I error rate control in all of the simulated combinations of variance ratios and standards. Therefore, the Type I error rates were completely controlled by all methods at both standards.

Sample size group 1 (n_1) = 40 and sample-size ratio (n_2/n_1) = 1.5. Table 11 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 1.5. Therefore, the sample size of the first group (i.e., n_1) was 40 and the sample size of the second group (i.e., n_2) was 60. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

When the standard was .05, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 11). In the case of the Good bootstrap method, the Type I error rates were relatively well controlled. It failed to control only when the variance ratio was 16. The pooled *t*-test method failed to control for Type I error rates on five occasions; therefore, it showed poor control over Type I error rates.

However, as expected, it controlled for Type I error rates when the variances were equal (i.e., the variance ratio was 1).

Table 10

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 40; sample-size ratio (n_2/n_1) was 1.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0430	0.0430	0.0555	0.0460	0.0440
1/4	0.0485	0.0505	0.0615	0.0530	0.0525
1/2	0.0465	0.0505	0.0570	0.0505	0.0505
1	0.0505	0.0550	0.0620	0.0545	0.0545
2	0.0450	0.0465	0.0530	0.0480	0.0475
4	0.0435	0.0455	0.0560	0.0480	0.0455
16	0.0505	0.0505	0.0575	0.0535	0.0510
			Standard .01		
1/16	0.0080	0.0080	0.0120	0.0105	0.0085
1/4	0.0090	0.0095	0.0130	0.0110	0.0100
1/2	0.0065	0.0080	0.0110	0.0075	0.0075
1	0.0070	0.0090	0.0125	0.0095	0.0095
2	0.0085	0.0115	0.0150	0.0110	0.0110
4	0.0085	0.0095	0.0130	0.0105	0.0105
16	0.0110	0.0105	0.0165	0.0115	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 11, when the standard was .01, the C&C, the E&T bootstrap, the Good bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The pooled *t*-test method failed to control for Type I error rates only when the variance ratios were 4 and 16. Therefore, it showed relatively good control over Type I error rates.

Table 11

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 40; sample-size ratio (n_2/n_1) was 1.5; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0500	0.0530	0.0570	0.0220*	0.0510
1/4	0.0540	0.0555	0.0595	0.0315*	0.0560
1/2	0.0455	0.0480	0.0535	0.0325*	0.0470
1	0.0470	0.0520	0.0585	0.0500	0.0520
2	0.0490	0.0495	0.0555	0.0650	0.0505
4	0.0470	0.0465	0.0560	0.0740*	0.0480
16	0.0595	0.0580	0.0710*	0.1100*	0.0605
			Standard .01		
1/16	0.0105	0.0110	0.0150	0.0040	0.0110
1/4	0.0080	0.0105	0.0130	0.0055	0.0105
1/2	0.0100	0.0120	0.0135	0.0065	0.0115
1	0.0115	0.0125	0.0150	0.0125	0.0130
2	0.0120	0.0120	0.0140	0.0165	0.0120
4	0.0095	0.0100	0.0150	0.0235*	0.0105
16	0.0090	0.0095	0.0155	0.0420*	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

Sample size group 1 (n_1) = 40 and sample-size ratio (n_2/n_1) = 3.0. Table 12 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 3.0. Therefore, the sample size of the first group (i.e., n_1) was 40 and the sample size of the second group (i.e., n_2) was 120. The results for all the variance ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented. As shown in Table 12, the C&C, the E&T bootstrap, the Good bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all

of the variance ratios, at both standards. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1), at both standards.

Table 12

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 40; sample-size ratio (n_2/n_1) was 3.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0430	0.0455	0.0505	0.0020*	0.0455
1/4	0.0455	0.0490	0.0535	0.0060*	0.0490
1/2	0.0505	0.0535	0.0585	0.0215*	0.0530
1	0.0420	0.0455	0.0545	0.0450	0.0460
2	0.0505	0.0500	0.0595	0.0945*	0.0505
4	0.0525	0.0535	0.0605	0.1420*	0.0530
16	0.0485	0.0480	0.0585	0.2175*	0.0485
			Standard .01		
1/16	0.0085	0.0090	0.0105	<0.0005*	0.0095
1/4	0.0090	0.0105	0.0120	<0.0005*	0.0105
1/2	0.0095	0.0110	0.0125	0.0025*	0.0105
1	0.0080	0.0090	0.0120	0.0090	0.0090
2	0.0085	0.0095	0.0145	0.0260*	0.0100
4	0.0110	0.0110	0.0150	0.0620*	0.0110
16	0.0130	0.0125	0.0165	0.1110*	0.0130

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

Sample size group 1 (n_1) = 40 and sample-size ratio (n_2/n_1) = 5.0. Table 13 contains the proportions of significant results when the null hypothesis ($\mu_1 = \mu_2 = 0$) was true, when the sample sizes ratio (i.e., n_2/n_1) was 5.0. Therefore, the sample size of the first group (i.e., n_1) was 40 and the sample size of the second group (i.e., n_2) was 200. The results for all the variance

ratios and the respective standards (i.e., the significance levels of the hypothesis tests of equality of the means) are presented.

Table 13

Proportions of significant results based on 2,000 replications of the simulated experiment when H_o ($\mu_1 = \mu_2 = 0$) was true; the standards were .05 and .01; n_1 was fixed to 40; sample-size ratio (n_2/n_1) was 5.0; var_2 was fixed to 1; and variance ratio was equal to var_1/var_2

Variance ratio	Methods				
	C&C	E&T	Good	Pooled	Satterthwaite
			Standard .05		
1/16	0.0495	0.0505	0.0530	<0.0005*	0.0510
1/4	0.0425	0.0425	0.0455	0.0010*	0.0430
1/2	0.0505	0.0510	0.0595	0.0130*	0.0515
1	0.0475	0.0485	0.0540	0.0445	0.0490
2	0.0480	0.0475	0.0575	0.1245*	0.0490
4	0.0545	0.0560	0.0635	0.2060*	0.0555
16	0.0525	0.0510	0.0590	0.3280*	0.0525
			Standard .01		
1/16	0.0060	0.0060	0.0085	<0.0005*	0.0065
1/4	0.0090	0.0095	0.0110	<0.0005*	0.0095
1/2	0.0090	0.0095	0.0110	0.0010*	0.0095
1	0.0095	0.0095	0.0120	0.0080	0.0100
2	0.0060	0.0060	0.0105	0.0350*	0.0060
4	0.0095	0.0090	0.0165	0.0945*	0.0095
16	0.0140	0.0130	0.0195*	0.1980*	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at $\alpha = .001$).

As shown in Table 13, when the standard was .05, the C&C, the E&T bootstrap, the Good bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1).

When the standard was .01, the C&C, the E&T bootstrap, and the Satterthwaite *t*-test methods showed complete Type I error rate control in all of the variance ratios (Table 13). In the case of the Good bootstrap method, the Type I error rates were relatively well controlled. It failed to control only when the variance ratio was 16. The pooled *t*-test method failed to control for Type I error rates on all occasions, except when the variances were equal (i.e., variance ratio was 1).

Overall summary of Type I error rates, n_1 = 40 and standard = .05. The C&C, E&T bootstrap, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. The Good bootstrap method failed to control for Type I error rates on only one occasion; therefore, it showed an almost complete control of Type I error rates.

The pooled *t*-test method showed complete Type I error rate control only when the sample sizes were equal (i.e., the sample-size ratio was 1.0). It showed poor control for Type I error rates when the sample-size ratio was 1.5 and failed to control in almost all of the variance ratios when the sample-size ratios were 3.0 and 5.0. As expected, it showed adequate control for Type I error rates when the variance ratios were 1, regardless of the sample-size ratios.

Overall summary of Type I error rates, n_1 = 40 and standard = .01. The C&C, E&T, and Satterthwaite *t*-test methods showed complete Type I error rate control in all of the simulated combinations of sample-size ratios and variance ratios. The Good bootstrap method failed to control for Type I error rates on one occasion; therefore, it showed almost complete control of Type I error rates.

The pooled *t*-test method showed complete Type I error rate control only when the sample sizes were equal (i.e., the sample-size ratio was 1.0). It showed relatively good control for Type I error rates when the sample-size ratio was 1.5, but failed to control in almost all of the

variance ratios when the sample-size ratios were 3.0 and 5.0. As expected, it showed adequate control for Type I error rates when the variance ratios were 1, regardless of the sample-size ratios.

Summary of the Type I error rates results. Among the five methods studied, only three (the C&C *t*-test, the nonparametric bootstrap method using the E&T approach, and the Satterthwaite approximate *t*-test) consistently yielded accurate *p*-values in a two-tailed hypothesis test, given each of the conditions of the present study. Note that only the Satterthwaite approximate *t*-test controlled for Type I error rates in all of the studied conditions. The C&C *t*-test failed to control in only two instances (Table 2). Similarly, the E&T bootstrap approach also failed to control in only two instances (Tables 2 and 3). However, in all cases in which the latter two methods failed, the Type I error rates were still close to the cut-off values for non-significant results (Table 1).

In the following section of this chapter, results are presented mostly for the power analysis of the three methods that consistently yielded accurate *p*-values in a two-tailed hypothesis test, given each of the conditions of the present study. That is, only the results from the power analysis of the C&C *t*-test, the E&T bootstrap, and the Satterthwaite *t*-test will be presented in most instances.

Power Analysis

A power study was conducted for the five methods under the same conditions evaluated for the Type I error rate study. However, given that only those methods that adequately control for Type I error rates should be used in practice, most of the present discussion is based only on the three methods discussed in the first part of this chapter (i.e., the Type I error rates study) that adequately controlled for Type I error rates most of the time. On the other hand, given that the

pooled *t*-test showed control for Type I error rates in cases when the sample sizes were equal or when the variances were equal, the results from this test are also included for those cases.

The power study was conducted using four equally spaced points (i.e., mean differences[15]) on the power curve. The four mean differences were 0.5, 1.0, 1.5, and 2.0. Also in this section, the variance ratios were grouped in three categories; small-variance ratios (i.e., var_1 = 1/16, 1/4, and 1/2), equal variances (i.e., $var_1 = var_2$), and large-variance ratios (i.e., var_1 = 2, 4, and 16). Note that the variance ratios are really defined by the variance of the first group (i.e., var_1) given that the variance of the second group (i.e., var_2) was fixed to 1.

Each graph contains a power curve of the theoretical power based on the corresponding simulated conditions and significance levels. These theoretical power curves were constructed based on the power values reported by the POWER procedure of the statistical software for the *t*-test using the Satterthwaite method. There were two reasons for including a theoretical power curve in each graph. First, the curve serves as a visual comparison with the power curves of the studied methods. Second, the curve is helpful as a confirmation of the internal validity of the results. That is, if the theoretical power curve overlaps or is very close to the power curve of the empirical estimation of power of the Satterthwaite *t*-test method, it provides a good confirmation of the validity of the simulation results.

Power results.[16] Among the three methods that consistently yielded accurate *p*-values for two-tailed hypothesis test (i.e., the C&C *t*-test, the E&T bootstrap approach, and the Satterthwaite approximate *t*-test), in almost all of the simulated conditions the Satterthwaite

[15] Although the statistical power analysis of a test is evaluated only when the H_0 is false (i.e., $\mu_1 \neq \mu_2$), the results when the Ho was true (i.e., no mean difference or $\mu_1 = \mu_2 = 0$) is also included in this study, as the starting mean difference for the power curves.

[16] In this section, the methods were ordered by the effectiveness with which they detect the means difference (i.e., power).

approximate *t*-test was slightly more powerful in detecting the mean differences at significance levels (i.e., standards). In some cases, the C&C *t*-test was slightly less powerful in detecting the mean differences, whereas in others the E&T bootstrap approach was slightly less powerful. However, in most instances the arithmetical differences in power were not large (i.e., the differences in power were negligible). Moreover, in many instances, the power curves of the three methods seem to be largely indistinguishable from each other. In the special cases where the variances were equal (e.g., variance ratio = 1), the pooled *t*-test method was the most powerful method for detecting the mean differences compared to the methods that adequately controlled for Type I error rates in most instances. When the sample sizes were equal (i.e., sample-size ratios were 1.0), most of the time the pooled *t*-test method was slightly more powerful than the other method of detecting the mean differences. In summary, in almost all the simulated conditions, one of the six patterns described above were seen. These patterns can be classified as those cases when Satterthwaite approximate *t*-test was the most powerful, those when the Cochran-Cox *t*-test was slightly less powerful, those when the Efron and Tibshirani bootstrap approach was slightly less powerful, those when the power curves were indistinguishable, and the two special cases when the pooled method was slightly more powerful (i.e., those cases when the variances were equal and those cases when the sample sizes were equal).

In the remainder of this section, a few selected power curves and their respective tables are shown to support the findings or six patterns described in the previous paragraph. It is important to clarify that although only few graphs and tables are presented here as examples of the power curve patterns observed, all of the graphs, as well as the tables with all of the power

results, are included in appendices A, B, and C. The graphs and tables in the appendices include the results of all methods, even if they failed to control for Type I error rates most of the time.

The Satterthwaite approximate t-test was the most powerful. In the case of unequal sample sizes ($n_1 \neq n_2$), when the sample size of the first group was relatively small (i.e., $n_1 = 10$), the sample size difference was also small (i.e., sample-size ratio of 1.5 or $n_2 = 15$), and when the lowest sample size (i.e., n_1) occurred with the highest variances (i.e., $var_2 = 2$, 4, and 16), at a significance level of .01, the power curves were relatively close. The Satterthwaite approximate *t*-test was the most powerful method in detecting the mean differences (Figure 1 and Table 14).

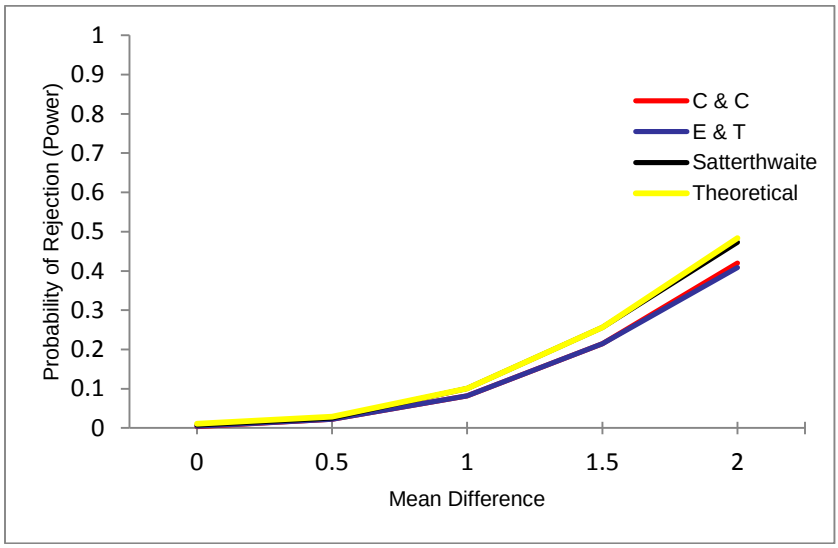

Figure 1. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 14

Power results[17] based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 1.5; and the variance ratio was 4; at a significance level (standard) of .01

Mean Difference	Methods		
	C&C	E&T	Satterthwaite
0.0	0.004	0.005	0.008
0.5	0.022	0.023	0.026
1.0	0.082	0.082	0.101
1.5	0.215	0.215	0.256
2.0	0.420	0.409	0.473

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Cochran-Cox t-test was slightly less powerful. In the case of unequal sample sizes ($n_1 \neq n_2$), when the sample size of the first group was relatively small (i.e., $n_1 = 10$), the sample size difference was also small (i.e., sample-size ratio of 1.5 or $n_2 = 15$), and when the lowest sample size (i.e., n_1) was combined with the lowest variances (i.e., $var_1 = 1/16, 1/4$, and $1/2$), at a significance level of .05, the power curves of most of the methods were basically indistinguishable. That is, most of the methods were very similar in detecting the mean differences (Figure 2 and Table 15). The C&C method showed less power, however.

[17] Note that the power results of the selected power tables shown in this subsection have been rounded to only three decimal places, to facilitate the interpretation of the results.

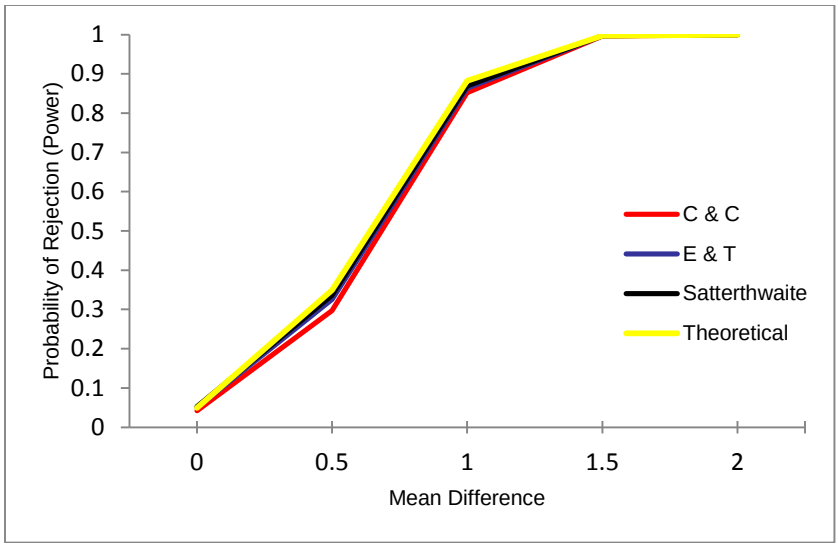

Figure 2. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 15

Power results based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 1.5; and the variance ratio was 1/4; at a significance level (standard) of .05

Mean Difference	Methods		
	C&C	E&T	Satterthwaite
0.0	0.042	0.051	0.053
0.5	0.298	0.326	0.336
1.0	0.853	0.864	0.870
1.5	0.997	0.997	0.998
2.0	1.000	1.000	1.000

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

The Efron and Tibshirani bootstrap approach was slightly less powerful. In the case of unequal sample sizes ($n_1 \neq n_2$), when the sample size of the first group was relatively small (i.e., $n_1 = 10$), the sample size difference was also small (i.e., sample-size ratio of 1.5 or $n_2 = 15$). When the lowest sample size (i.e., n_1) was combined with the lowest variances (i.e., $var_1 = 1/16$, 1/4, and 1/2), at a significance level of .01, the power curves of most of the methods were

basically indistinguishable. That is, most of the methods were very similar in detecting the mean differences (Figure 3 and Table 16). The E&T method showed slightly less power, however.

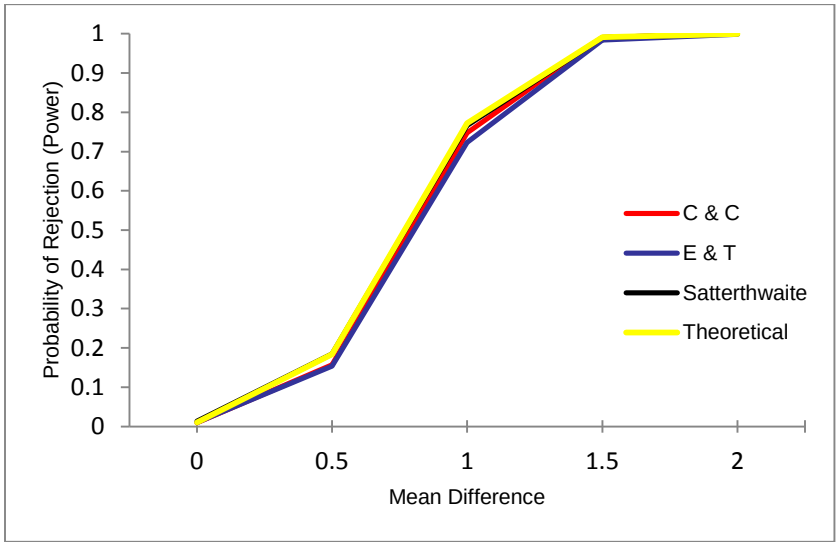

Figure 3. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 16

Power results based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 1.5; and the variance ratio was 1/16; at a significance level (standard) of .01

Mean Difference	Methods		
	C&C	E&T	Satterthwaite
0.0	0.010	0.011	0.014
0.5	0.157	0.154	0.185
1.0	0.748	0.724	0.767
1.5	0.991	0.985	0.991
2.0	1.000	0.999	1.000

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Power curves indistinguishable. In the case of unequal sample sizes ($n_1 \neq n_2$), when the sample size of the first group was relatively small (i.e., $n_1 = 10$), the sample size difference was large (i.e., sample-size ratio of 5.0 or $n_2 = 50$), and when the lowest sample size (i.e., n_1) was

combined with the lowest variances (i.e., var_1 = 1/16, 1/4, and 1/2), at a significance level of .05, the power curves were basically indistinguishable (Figure 4 and Table 17).

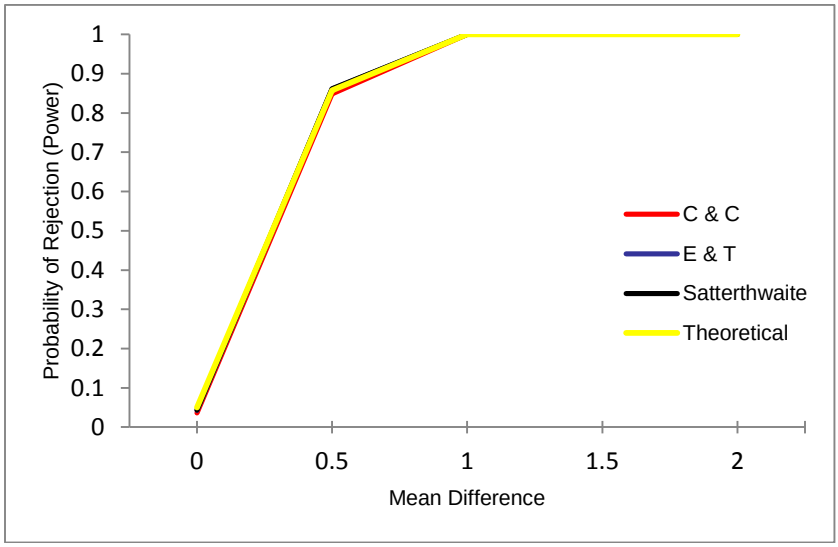

Figure 4. Power curves for unequal group sample sizes when n_1 = 10, n_2 = 50, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 17

Power results based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 5.0; and the variance ratio was 1/16; at a significance level (standard) of .05

Mean Difference	Methods		
	C&C	E&T	Satterthwaite
0.0	0.036	0.043	0.044
0.5	0.849	0.861	0.862
1.0	1.000	1.000	1.000
1.5	1.000	1.000	1.000
2.0	1.000	1.000	1.000

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Equal variances. As we can recall, when the variance ratio was 1 (i.e., $var_1 = var_2$), the pooled *t*-test consistently showed adequate control for Type I error rates (i.e., it did not fail to control). In the case of unequal sample sizes ($n_1 \neq n_2$), when the sample size of the first group

was relatively small (i.e., n_1 = 10), the sample size difference was also small (i.e., sample-size ratio of 1.5 or n_2 = 15), at a significance level of .01, the power curves were very close but the most powerful method was the pooled method; the C&C method was the least powerful (Figure 5 and Table 18).

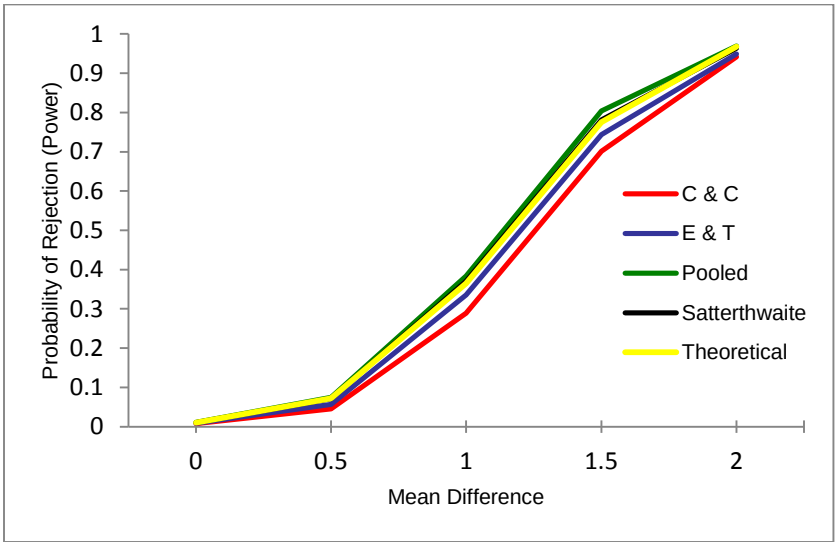

Figure 5. Power curves for unequal group sample sizes when n_1 = 10, n_2 = 15, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 18

Power results based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 1.5; and the variance ratio was 1; at a significance level (standard) of .01

Mean Difference	Methods			
	C&C	E&T	Pooled	Satterthwaite
0.0	0.008	0.009	0.011	0.010
0.5	0.046	0.057	0.075	0.072
1.0	0.289	0.335	0.384	0.371
1.5	0.702	0.744	0.804	0.782
2.0	0.942	0.949	0.969	0.964

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Equal sample sizes. As we can recall, when the sample size ratio was 1.0 (i.e., $n_1 = n_2$), the pooled *t*-test consistently showed adequate control for Type I error rates. In the case of equal

sample sizes ($n_1 = n_2 = 10$), when the variance ratio was 1/16, at a significance level of .05, the power curves were almost indistinguishable but the pooled method was slightly more powerful than the others (Figure 6 and Table 19).

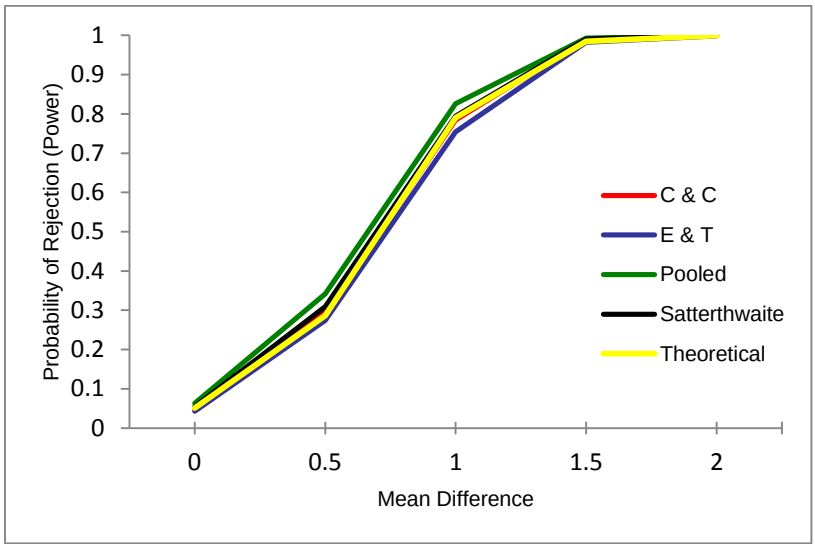

Figure 6. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.

Table 19

Power results based on 2,000 replications of the simulated experiment when n_1 was fixed to 10; the sample-size ratio (n_2/n_1) was 1.0; and the variance ratio was 1/16; at a significance level (standard) of .05

Mean Difference	Methods			
	C&C	E&T	Pooled	Satterthwaite
0.0	0.048	0.043	0.064	0.055
0.5	0.295	0.275	0.342	0.310
1.0	0.785	0.755	0.826	0.795
1.5	0.990	0.983	0.993	0.991
2.0	1.000	0.999	1.000	1.000

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Summary of the power analysis results. The six sets of graphs and tables presented above are representative of all the possible results of the power analysis. As previously mentioned, in most instances the differences in power were mostly negligible and even in many

of those instances, the power curves of the three methods mostly seem to be indistinguishable from each other. However, when the variances were equal (e.g., variance ratio = 1), the pooled *t*-test was the most powerful method for detecting the mean differences. Similarly, when the sample sizes were equal (i.e., sample-size ratios were 1.0) the pooled *t*-test method was slightly more powerful than the other method in detecting the mean differences, most of the time.

The next chapter (i.e., Chapter 5), consist of a discussion about the results presented here, including some of the implications. Also, several recommendations for practice are included, followed by some of the limitations of the present study. Chapter 5 ends with the conclusions of this study as well as some suggestions for future research.

Chapter 5: Discussion

The main purpose of this study was to compare and contrast the Type I error rate performance and statistical power of five approaches (tests or methods) to the Behrens–Fisher problem under several simulated conditions. The methods studied were the Cochran and Cox *t*-test assuming unequal variances, a non-parametric bootstrapping method using the Efron and Tibshirani (1993) approach, a non-parametric bootstrapping method using the Good (2005) approach, a pooled *t*-test or classical *t*-test, and the approximate *t*-test with a Welch–Satterthwaite correction.

Summary of Research Problem and Methods Used

It is generally accepted that the pooled *t*-test, also known as the Student's *t*-test or classical *t*-test, is the uniformly most powerful unbiased (UMPU) test for the equality of two independent population means if the assumptions of normality and HOV are not violated (Olejnik & Luh, 1994). However, if the variances are heterogeneous, a condition known as heteroscedasticity, the Type I error rate (i.e., the significance level of the test) is no longer stable (Olejnik & Luh, 1994). That is, the significance level of the *t*-test could be actually greater or lower than the pre-assigned level (i.e., the significance level set by the researcher; nominal α). In

other words, the results obtained with this test may not be valid or accurate if all assumptions of the test, including the HOV, are not met. In the statistical literature, the violation of the unequal variances assumption when comparing the difference of two independent means has been called the Behrens–Fisher problem (Howell, 2002; Kim & Cohen, 1998; Pesarin, 1995) in reference to the first two statisticians who are known to have dealt with it. This problem is still "of theoretical interest in statistics because there is no exact solution to such an apparently simple problem" (van Belle et al., 2004, p. 139).

Numerous methods or approaches have been proposed to resolve the Behrens–Fisher problem (Aspin & Welch, 1949; Efron & Tibshirani, 1993; Fisher, 1935; Hyslop & Lupinacci, 2003; Lee & Gurland, 1975; Satterthwaite, 1946; Sawilowsky, 2002; Scheffé, 1970; Wang & Chow, 2002; Welch, 1947). Some of the approaches are parametric-based and others are nonparametric-based. Two of the most commonly used parametric-based approaches are the Welch–Satterthwaite solution and the Cochran–Cox *t*-test. Two of the modern proposed nonparametric solutions are the Good bootstrapping approach and the Efron and Tibshirani bootstrapping approach. To my knowledge, no other study has been conducted to compare and contrast these five approaches (tests or methods), including the pooled *t*-test, in terms of Type I error rate performance and statistical power.

To conduct the present Monte Carlo study, several samples from a normal population with mean 0 and variance 1 were generated given different sample sizes, means, and variance conditions to form each of the two independent groups. The study was conducted, based upon the simulated data, by applying different dimensions (sample sizes, sample variances, and mean differences). The resulting data were used to conduct simulated experiments for Type I error

rates and power analyses for the five selected methods. The final step consisted of comparing and contrasting the results of the Type I error rate simulations and power analyses.

Overall Summary of Results

Type I error rates. Of the five methods studied, under the assumption of heterogeneity of variances (i.e., the Behrens–Fisher problem), only three (listed in alphabetical order) consistently yielded accurate *p*-values for a two-tailed hypothesis test given each of the conditions of this study: the Cochran–Cox *t*-test, the nonparametric bootstrap method using the Efron–Tibshirani approach, and the Welch–Satterthwaite approximate *t*-test. Of these, only the Welch–Satterthwaite approximate *t*-test completely controlled the Type I error rates in all of the studied conditions. The Cochran–Cox *t*-test failed to control in only two instances (Table 2). Similarly, the Efron–Tibshirani bootstrap approach also failed to control in only two instances (Tables 2 and 6). However, in all of the cases in which these two methods failed, the Type I error rates were still close to the cut-off value for non-significant results. On the other hand, when the sample sizes were equal, the pooled *t*-test also yielded accurate *p*-values in almost all instances.

Power analysis. Among the three methods that consistently yielded accurate *p*-values for a two-tailed hypothesis test (i.e., the Cochran–Cox *t*-test, the nonparametric bootstrap method using the Efron–Tibshirani approach, and the Welch–Satterthwaite approximate *t*-test), in almost all of the simulated conditions the Welch–Satterthwaite approximate *t*-test was the most powerful test in detecting the mean differences, at both significance levels, when the variances were heterogeneous (i.e., the Behrens–Fisher problem). In some cases the Cochran–Cox *t*-test was slightly less powerful in detecting the mean differences, whereas in others the Efron–Tibshirani bootstrap approach was slightly less powerful. In most instances, however, the arithmetical differences in power were not large; therefore, on most occasions the power curves of these three methods seemed to be indistinguishable.

In the special cases when the sample sizes were equal and the pooled *t*-test method controlled for Type I error rates, it was the most powerful method in detecting the mean differences in comparison to the methods that adequately controlled for Type I error rates most of the time. Similarly, in other special cases when the variances were equal, the pooled *t*-test was the most powerful method in detecting the mean differences in comparison to the methods that adequately controlled for Type I error rates most of the time. However in the latter special cases, the situation did not involve a Behrens–Fisher problem because the variances were homogeneous.

Interpretation of Results

This section is divided into two parts according to the two main purposes of the present study. As in previous divided sections, the first part is about the Type I error rates and the second part is about the power analysis. Please note that the main focus of the present study has been to evaluate and compare the Type I error rates and the power of five methods that have been proposed in the literature as alternatives for hypothesis testing of a mean difference between two samples when the variances are different (i.e., the Behrens–Fisher problem).

Type I error rates. The first research question of the present study was: Can certain nonparametric bootstrap methods (i.e., a nonparametric bootstrapping method using the Good approach and a nonparametric bootstrapping method using the Efron–Tibshirani approach), the Welch–Satterthwaite approximate *t*-test, the classical *t*-test, and the Cochran–Cox *t*-test, yield accurate *p*-values for a two-tailed hypothesis test, given each of the conditions of the present study? As expected, based on my literature review as well as the vast experience conveyed by previous studies, I mostly obtained accurate results about Type I error rates from the Cochran–Cox *t*-test assuming unequal variances, as well as from the approximate *t*-test with Welch–Satterthwaite correction (i.e., both of these *t*-tests yielded accurate *p*-values). The accurate results

were for two-tailed hypothesis tests of mean difference given almost all of the simulated conditions. Also as expected, the classical *t*-test (i.e., the pooled *t*-test) mostly failed to yield accurate *p*-values except in two kinds of situations. The first was when the samples were equal (i.e., sample-size ratios were 1.0), regardless of the variance ratio, except for one occasion when $n_1 = n_2 = 10$. The second kind of situation in which the pooled *t*-test yielded accurate *p*-values was all occasions when the variance ratios were 1 (i.e., $var_1 = var_2$), regardless of the sample-size ratio. As previously mentioned, the classical *t*-test is the uniformly most powerful unbiased (UMPU) test for the equality of two independent population means—but only if the assumption homogeneity of variance (i.e., $var_1 = var_2$) is not violated. If the variances are heterogeneous ($var_1 \neq var_2$), the Type I error rate (i.e., the significance level of the test) is no longer stable and may not be valid or accurate (Chapter 4). Therefore, the results of the present study for the classical *t*-test (i.e., pooled *t*-test) are consistent with the literature in cases when the homogeneity of variance assumption was not violated, and also in cases where the samples were equal (i.e., sample-size ratio was 1.0). Thus the results of the Cochran–Cox *t*-test assuming unequal variances, the approximate *t*-test with Welch–Satterthwaite correction, and the classical *t*-test (i.e., assumptions of equal variance), give internal validity to this study. That is, the results of those methods were mostly as expected and in concordance with results previously reported and published.

With respect to the two nonparametric bootstrapping methods considered in the present study, only the method using the Efron–Tibshirani (1993) approach yielded accurate *p*-values values in almost all instances; the Good (2005) approach did not. As previously mentioned, based on the literature reviewed for the present study, these two particular bootstrapping approaches were specifically designed or proposed for comparisons of the differences between

two means under the heterogeneity of variances. However, to my knowledge, these two approaches have not been extensively employed in Monte Carlo studies that compare different approaches to or methods of resolving the Behrens–Fisher problem, in contrast to the vast literature about studies on the other most common methods, such as the Cochran–Cox *t*-test assuming unequal variances and the approximate *t*-test with Welch–Satterthwaite correction. Nonetheless, given the specific design or purpose of these two bootstrap approaches, it was a surprise that one of them, the Good (2005) bootstrapping approach, failed to yield accurate *p*-values in so many instances.

As discussed in Chapter 4, the Good (2005) bootstrapping approach failed in most instances when the sample sizes of the first group were relatively small (i.e., $n_1 = 10$) and several cases when the size of the first group was 25. However, it yielded accurate *p*-values in almost all instances when the size of the first group was relatively large (i.e., $n_1 = 25$ or greater). Therefore, it seems that the accuracy of this approach, in terms of *p*-values, increases as the total sample size (i.e., $n_1 + n_2$) also increases. That is, the larger the total sample size, the more accurate the *p*-values of the Good (2005) bootstrapping approach, to the point that when the sample size of the first group was relatively large ($n_1 = 40$), the Good bootstrapping method showed almost complete control over Type I error rates under the simulated conditions. The Good (2005) bootstrap approach was not able to yield accurate *p*-values when total sample size is as small as 60 or less, especially when the sample size of the first group (n_1) is 10. Good (2013) himself acknowledged as much when wrote, "Warning: The bootstrap is not recommended for use with small samples…" (p. 86). Although he may have been making a general recommendation about using bootstrap methods, the results of the present study indicate that this warning is directly applicable to his approach.

By contrast, the other bootstrap approach for hypothesis testing evaluated in the present study, the Efron–Tibshirani (1993) approach, yielded accurate *p*-values in almost all instances, including basically all of the cases in which sample sizes were relatively small. This method only failed to yield to adequate control for Type I error rate in two instances: at a standard (i.e., significance level) of 0.05, when $n_1 = n_2 = 10$ and the variance ratio ($variance_1/variance_2$) was 4; and at the same standard when $n_1 = 10$, $n_2 = 30$ and the variance ratio ($variance_1/variance_2$) was 16. Therefore, given the simulated conditions of the present study, it seems that the Efron–Tibshirani (1993) approach, contrary to what was observed in the present study for the Good (2005) approach, works not only for relatively large sample sizes but also for total sample sizes as small as 20 when $n_1 = n_2 = 10$.

Power analysis. The second research question of the present study was: Is the hypothesis test based on the proposed nonparametric bootstrap methods, the classical *t*-test, and the Cochran–Cox *t*-test, more powerful than the Welch–Satterthwaite approximate *t*-test, given each of the conditions of this study? The clear answer to this question is no. The two methods that consistently yield accurate *p*-values for two-tailed hypothesis test, the Cochran–Cox *t*-test and Efron–Tibshirani bootstrap approach, in general do not seem to be more powerful than the Welch–Satterthwaite approximate *t*-test method. On the contrary, although most of the time the power curves of these three methods or approaches were very close, on many occasions one of the two methods (i.e., the Cochran–Cox *t*-test and the Efron–Tibshirani bootstrap approach) was slightly less powerful than the Welch–Satterthwaite approximate *t*-test method.

On the other hand, in the special cases when the sample sizes were equal or when the variances were equal, the pooled *t*-test (i.e., classical *t*-test) was the most powerful method in detecting the mean differences most of the time, as previously discussed. This conclusion, with

respect to cases when the variances were equal, does not support the recommendation given by Heiser (2006). As mentioned in chapter 2 of this dissertation, Heiser (2006) recommended that "if the assumption of normality is valid, then the best method is the *v*-test [using the Welch adjustment for the degrees of freedom]…for all tests on the difference in means, regardless if the variances are equal or unequal" (p. 563). However, the results obtained in the present study, with respect to the power of the Welch–Satterthwaite approximate *t*-test do not support that recommendation. As mentioned before, in the present study when the variances were equal, the pooled *t*-test was the most powerful method in detecting the mean differences. Therefore, based on these results it does not seem reasonable to prefer the use of the Welch–Satterthwaite approximate *t*-test instead of the pooled *t*-test (i.e., classical *t*-test) to conduct hypothesis testing of the mean differences of two groups when the variances are equal.

Implications of the Results and Recommendation for Practice

The present study has shown that when the performance of the five methods are compared, in cases that contain a Behrens–Fisher problem situation, given each of the simulated conditions of the present study, only three methods consistently control for Type I error rates. Those three methods were the Cochran–Cox *t*-test, the nonparametric bootstrapping approach using the Efron–Tibshirani (1993) method, and the Welch–Satterthwaite approximate *t*-test. In other words, only these three methods yielded accurate *p*-values for a two-tailed hypothesis test, most of the time. Among them, only the Welch–Satterthwaite approximate *t*-test showed complete control for Type I error rates in any of the simulated conditions. The Cochran–Cox *t*-test and the nonparametric bootstrapping approach using the Efron–Tibshirani (1993) method failed to control in only two instances, both at a significance level (α) of 0.05. Therefore, it could

be said that the Type I error rate control demonstrated by these two methods is generally acceptable.

As is known, the Welch–Satterthwaite approximate *t*-test is a default test available in most, if not all, modern statistical software, for use when the variances of two samples are unknown and may be unequal (i.e., in cases that contain the Behrens–Fisher problem). Along with the classical *t*-test, it is perhaps one of the most common tests conducted by statisticians and researchers as they compare the equality of two means. The present study has shown that, among the methods or approaches applied to the Behrens–Fisher problem (i.e., those methods that in this study consistently yielded accurate *p*-values for two-tailed hypothesis test), the Welch–Satterthwaite approximate *t*-test is slightly more powerful in detecting the means difference between two samples. The present study has presented no compelling evidence, however, to indicate that among the five methods or approaches evaluated, a method other than the Welch–Satterthwaite approximate *t*-test provides a better alternative based on its simulated conditions, except when the sample sizes are equal. Given all of the evidence presented above, of the five methods evaluated in the present study, I recommend use of the Welch–Satterthwaite approximate *t*-test in cases when the samples have been obtained from a normally distributed population, when the sizes of these samples are not equal, and when there is uncertainty that the variances of the samples are equal. On the other hand, when sample sizes are equal or when the variances are equal, I recommend use of the pooled (classical) *t*-test.

Limitations of the Present Study

As with any research study, this one is not exempt from limitations. The first is that the present study was based on simulated data given several parameters. Therefore, its results in terms of Type I error rates and power analyses are definitively valid only for cases in which the

combination of conditions is similar to those simulated herein. However, the combination of conditions within the present study covered a relatively large variety of circumstances, similar to those that can be found in practice (i.e., while conducting statistical analyses with real data). Of course, there is no need to suspect that close but somehow different combinations of conditions than those simulated in the present study would give completely different results. However, at the same time there is no certainty that if a different combination of conditions was to be simulated or observed in a real data, the results would be similar to those observed herein.

Another limitation is that the simulations were conducted based on only 2,000 replications for each method. Therefore, although I suspect that the results would not be significantly different if more replications had been conducted (e.g., 5,000 or 10,000), I cannot declare that the results would imitate the ones observed herein. Early in the simulation phase of the present study, I discovered (among other issues) that the available computer resources, in terms of available memory space and allocated hours of use, were insufficient to conduct my simulations with numbers of replications higher than 2,000.

Last, it must be emphasized that the present results are based on normal data only. That is, the first condition of all the simulated data in the present study was normality. Therefore, the results of the present study apply only to data normally distributed, given the other combinations of conditions. If the data simulated were not normally distributed, the results as well as the recommendations for practice could be totally different.

Conclusions

When conducting a hypothesis testing of the equality of two means (i.e., comparisons of the mean difference of two samples) obtained from normal distributions whose variances are unknown and possibly different (i.e., a Behrens–Fisher problem), the Welch–Satterthwaite

approximate *t*-test showed an excellent control for Type I error rates, in contrast to the pooled *t*-test[18], the Cochran–Cox *t*-test, and the two bootstrap methods (a nonparametric bootstrapping method using the Good approach and a nonparametric bootstrapping method using the Efron–Tibshirani approach). Similarly, the Welch–Satterthwaite approximate *t*-test also showed in general the best power in detecting a mean difference among the three methods studied that consistently yield accurate *p*-values for two-tailed hypothesis test (i.e., the Cochran–Cox *t*-test, the Efron–Tibshirani bootstrap and the Welch–Satterthwaite approximate *t*-test). Therefore, for practice, there is no compelling evidence obtained from this study to suggest that among the evaluated methods or approaches, a method other than the Welch–Satterthwaite approximate *t*-test is a better alternative for resolving the Behrens–Fisher problem, except when the sample sizes are equal[19].

Suggestions for Future Research

As mentioned above, the results of the present study apply only to normally distributed data, given the other combinations of conditions. However, in many research situations, the data available for analysis are not normally distributed. Therefore, an extension of the present research should incorporate the non-normality aspect of the data in the simulations for future studies.

On the other hand, in the present study, only one type of bootstrap, the percentile, was considered when applying the modified version of the Good bootstrap approach or the Efron–Tibshirani bootstrap approach. However, there are several other types of bootstrap that could be

[18] As discussed before, in the special cases when the sample sizes were equal, the pooled *t*-test controlled the Type I error rates in almost all occasions.
[19] In cases when the sample sizes are equal, the pooled *t*-test seems to be a slightly better alternative than the Welch–Satterthwaite approximate *t*-test.

used. For future studies, other types of bootstrap (e.g., the parametric) can be used to confirm or reject the current findings of the present study with respect to Type I error rates and Power. Similarly, other modified versions to the Good bootstrap or to the Efron–Tibshirani bootstrap approaches could be studied, as well as other kinds of bootstrap approaches not considered in the present study.

APPENDIX A: TYPE I ERROR RATE TABLES, POWER TABLES, AND POWER CURVES, WHEN THE SAMPLE SIZE OF GROUP 1 (n_1) WAS 10

Sample-size Ratio was 1.0 (i.e., Equal Sample Size or $n_1 = n_2 = 10$)

Table A1

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0475	0.0085
E & T	0.0430	0.0080
Good	0.0980*	0.0370*
Pooled	0.0635	0.0185*
Satterthwaite	0.0545	0.0125

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A2

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0415	0.0065
E & T	0.0425	0.0080
Good	0.0800*	0.0320*
Pooled	0.0525	0.0110
Satterthwaite	0.0490	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A3

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0460	0.0040
E & T	0.0550	0.0065
Good	0.0920*	0.0330*
Pooled	0.0595	0.0125
Satterthwaite	0.0580	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A4

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0365	0.0040
E & T	0.0430	0.0080
Good	0.0725*	0.0295*
Pooled	0.0465	0.0110
Satterthwaite	0.0455	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A5

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0320*	0.0040
E & T	0.0360	0.0075
Good	0.0715*	0.0245*
Pooled	0.0410	0.0095
Satterthwaite	0.0380	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A6

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0295*	0.0060
E & T	0.0335*	0.0055
Good	0.0720*	0.0215*
Pooled	0.0435	0.0105
Satterthwaite	0.0395	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A7

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0425	0.0070
E & T	0.0370	0.0060
Good	0.0870*	0.0330*
Pooled	0.0540	0.0135
Satterthwaite	0.0465	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A8

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0475	0.0415	0.0460	0.0365	0.0320	0.0295	0.0425
	0.5	0.2950	0.2360	0.1815	0.1465	0.1085	0.0825	0.0575
	1.0	0.7845	0.7250	0.6225	0.4825	0.3675	0.2315	0.0995
	1.5	0.9895	0.9675	0.9285	0.8580	0.6720	0.4770	0.1705
	2.0	1.0000	1.0000	0.9965	0.9775	0.9115	0.6945	0.2760
Efron & Tibshirani	0.0	0.0430	0.0425	0.0550	0.0430	0.0360	0.0335	0.0370
	0.5	0.2745	0.2435	0.2035	0.1685	0.1225	0.0900	0.0520
	1.0	0.7545	0.7295	0.6495	0.5140	0.3935	0.2455	0.0895
	1.5	0.9825	0.9640	0.9385	0.8750	0.6995	0.4830	0.1570
	2.0	0.9985	0.9995	0.9965	0.9805	0.9210	0.7000	0.2555
Good	0.0	0.0980	0.0800	0.0920	0.0725	0.0715	0.0720	0.0870
	0.5	0.4120	0.3640	0.2950	0.2465	0.1885	0.1545	0.1100
	1.0	0.8790	0.8280	0.7495	0.6390	0.5010	0.3490	0.1725
	1.5	0.9970	0.9895	0.9680	0.9325	0.8085	0.6140	0.2710
	2.0	1.0000	1.0000	0.9995	0.9925	0.9600	0.8170	0.3915
Pooled	0.0	0.0635	0.0525	0.0595	0.0465	0.0410	0.0435	0.0540
	0.5	0.3420	0.2795	0.2265	0.1790	0.1370	0.1070	0.0770
	1.0	0.8260	0.7745	0.6750	0.5365	0.4150	0.2795	0.1245
	1.5	0.9930	0.9810	0.9470	0.8875	0.7310	0.5340	0.2145
	2.0	1.0000	1.0000	0.9975	0.9865	0.9370	0.7515	0.3200
Satterthwaite	0.0	0.0545	0.0490	0.0580	0.0455	0.0380	0.0395	0.0465
	0.5	0.3095	0.2630	0.2220	0.1770	0.1305	0.1005	0.0630
	1.0	0.7945	0.7570	0.6680	0.5330	0.4055	0.2640	0.1060
	1.5	0.9910	0.9745	0.9440	0.8840	0.7195	0.5135	0.1815
	2.0	1.0000	1.0000	0.9975	0.9860	0.9320	0.7340	0.2855

Table A9

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 10$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0085	0.0065	0.0040	0.0040	0.0040	0.0060	0.0070
	0.5	0.0915	0.0580	0.0415	0.0365	0.0270	0.0145	0.0110
	1.0	0.4775	0.3895	0.3030	0.1965	0.1075	0.0615	0.0230
	1.5	0.8800	0.8120	0.7160	0.5590	0.3535	0.2030	0.0490
	2.0	0.9880	0.9830	0.9550	0.8575	0.6775	0.3580	0.0900
Efron & Tibshirani	0.0	0.0080	0.0080	0.0065	0.0080	0.0075	0.0055	0.0060
	0.5	0.0755	0.0765	0.0570	0.0525	0.0335	0.0195	0.0100
	1.0	0.4045	0.3970	0.3550	0.2425	0.1355	0.0685	0.0165
	1.5	0.8055	0.8070	0.7585	0.6260	0.3855	0.2100	0.0395
	2.0	0.9585	0.9715	0.9545	0.8970	0.7220	0.3685	0.0735
Good	0.0	0.0370	0.0320	0.0330	0.0295	0.0245	0.0215	0.0330
	0.5	0.2490	0.1940	0.1460	0.1175	0.0935	0.0665	0.0435
	1.0	0.7495	0.6715	0.5600	0.4220	0.3110	0.1935	0.0840
	1.5	0.9770	0.9580	0.9005	0.8195	0.6110	0.4330	0.1415
	2.0	0.9995	0.9990	0.9940	0.9620	0.8855	0.6365	0.2410
Pooled	0.0	0.0185	0.0110	0.0125	0.0110	0.0095	0.0105	0.0135
	0.5	0.1405	0.1085	0.0790	0.0625	0.0410	0.0320	0.0180
	1.0	0.5900	0.5105	0.4275	0.2935	0.1790	0.1045	0.0440
	1.5	0.9380	0.8940	0.8205	0.6820	0.4620	0.2810	0.0835
	2.0	0.9950	0.9970	0.9770	0.9235	0.7910	0.4780	0.1375
Satterthwaite	0.0	0.0125	0.0105	0.0105	0.0110	0.0090	0.0090	0.0075
	0.5	0.1005	0.0955	0.0730	0.0595	0.0375	0.0290	0.0120
	1.0	0.5010	0.4610	0.4070	0.2840	0.1640	0.0910	0.0300
	1.5	0.8940	0.8625	0.8040	0.6730	0.4370	0.2565	0.0575
	2.0	0.9900	0.9915	0.9755	0.9200	0.7780	0.4320	0.1055

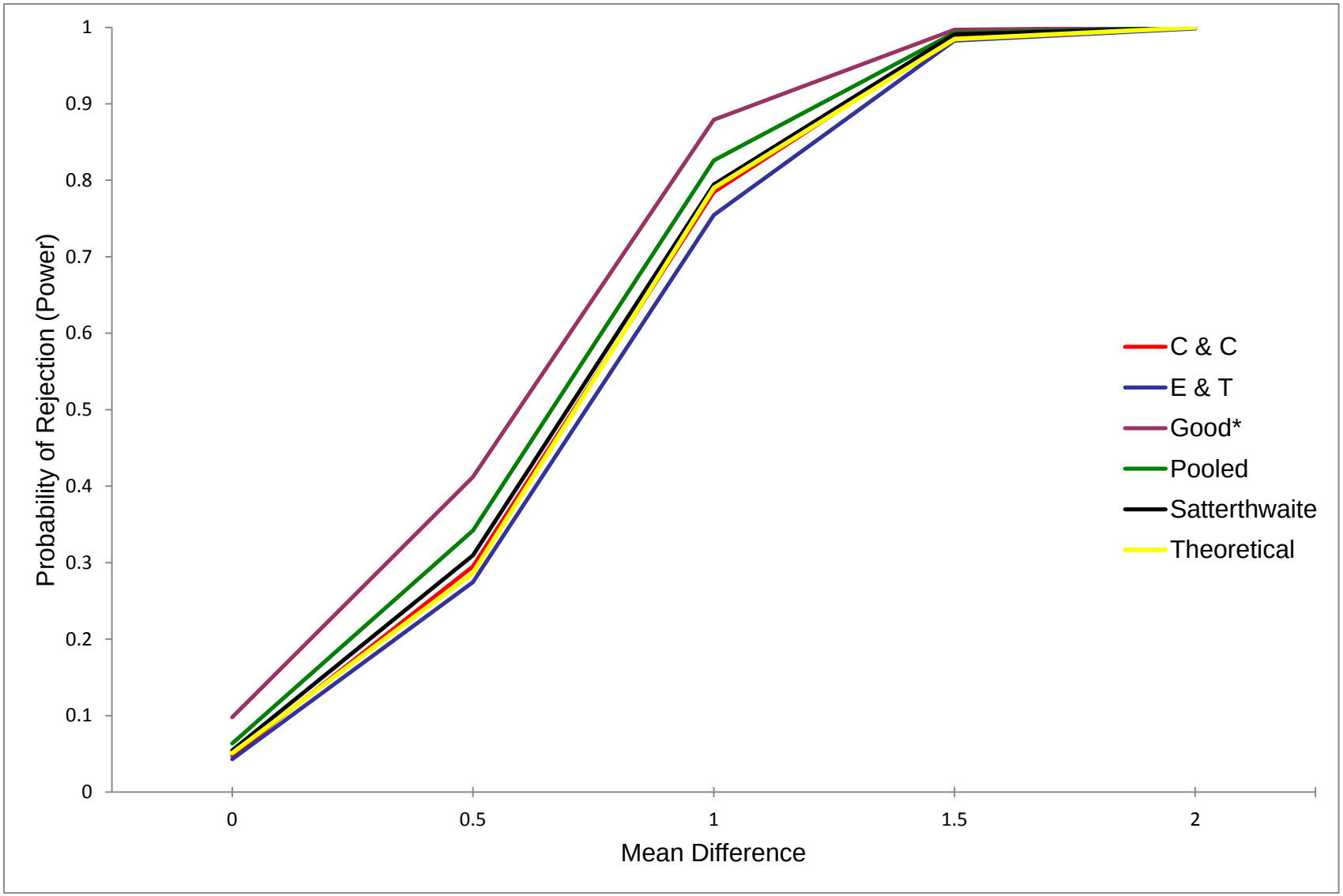

Figure A1. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

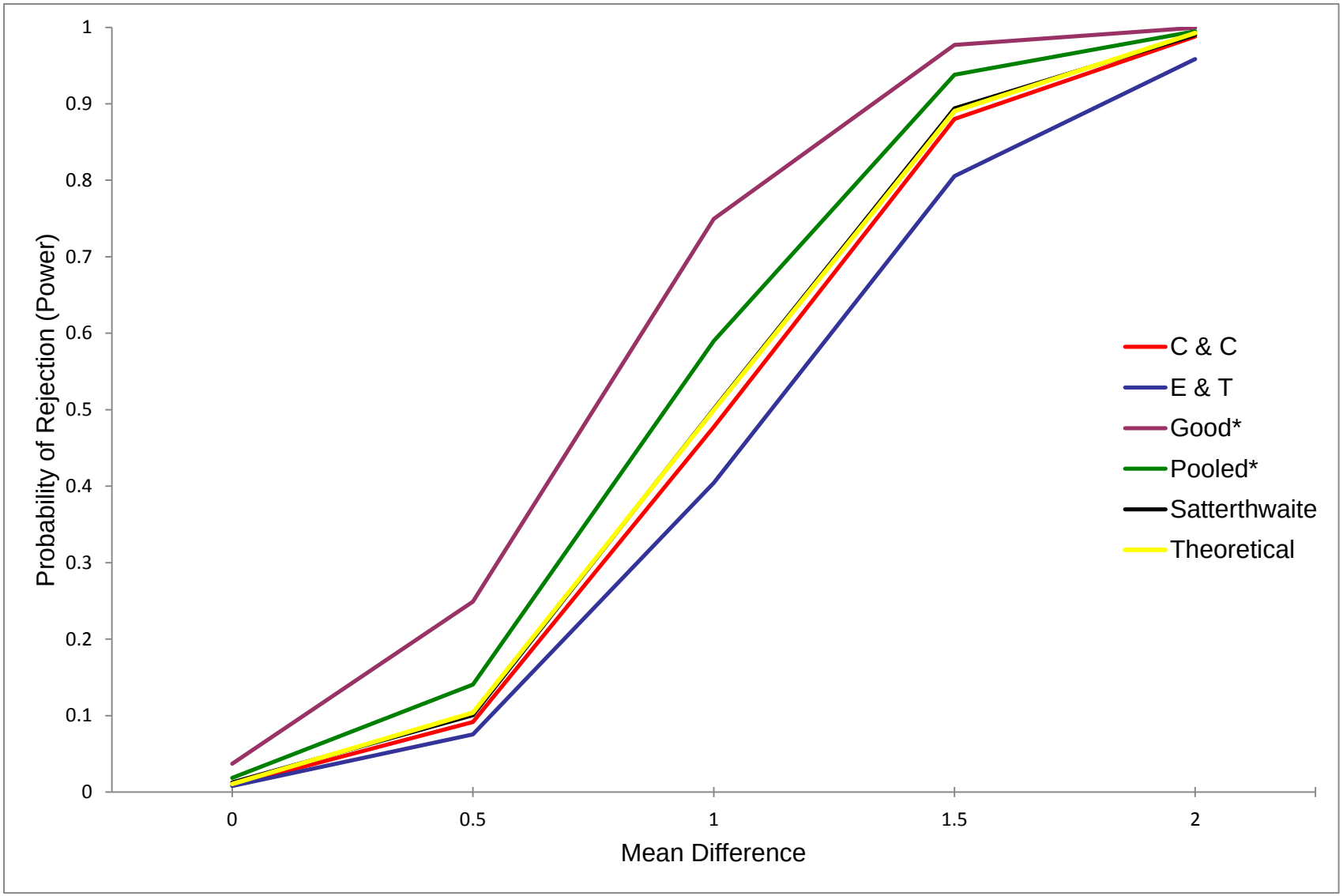

Figure A2. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

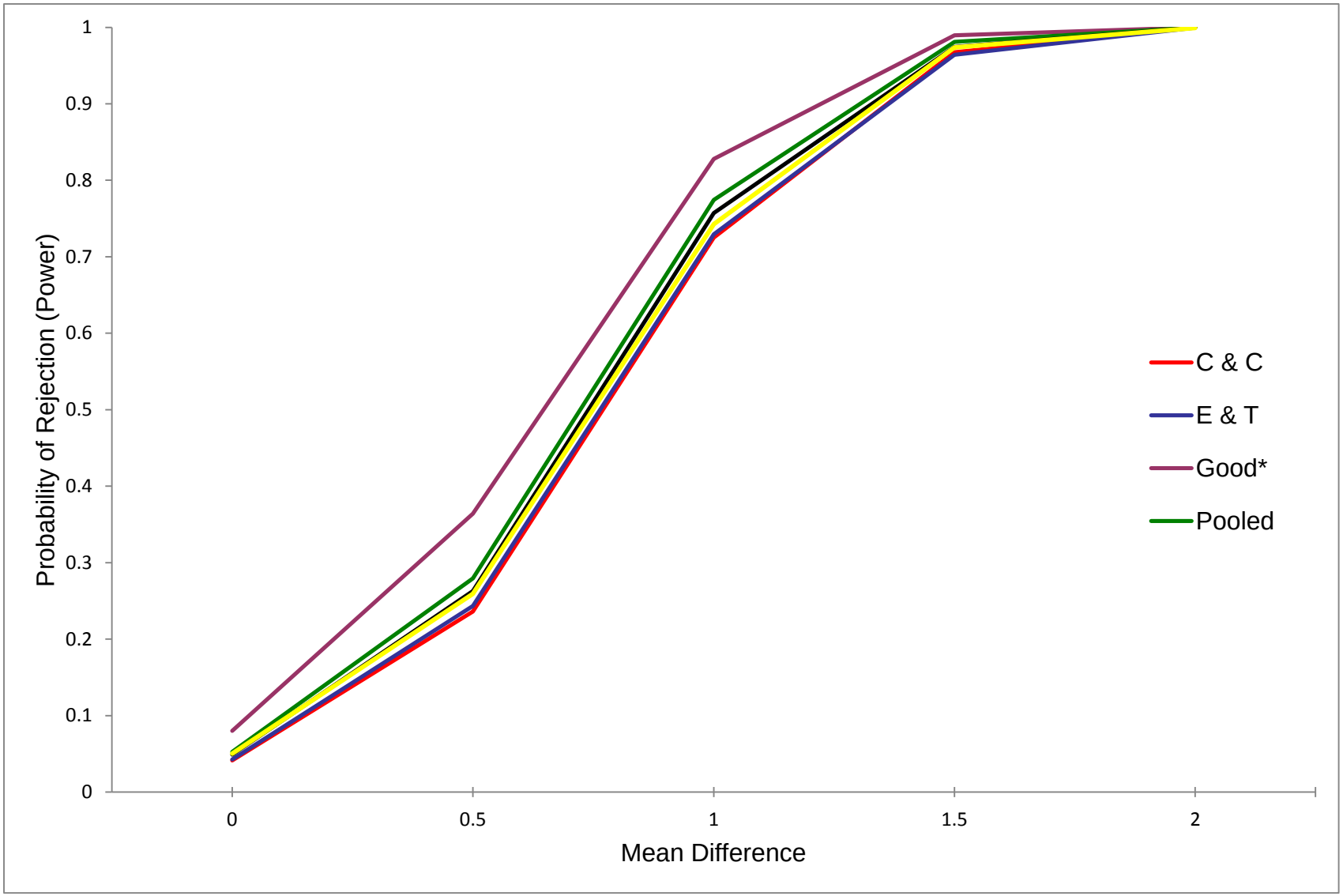

Figure A3. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

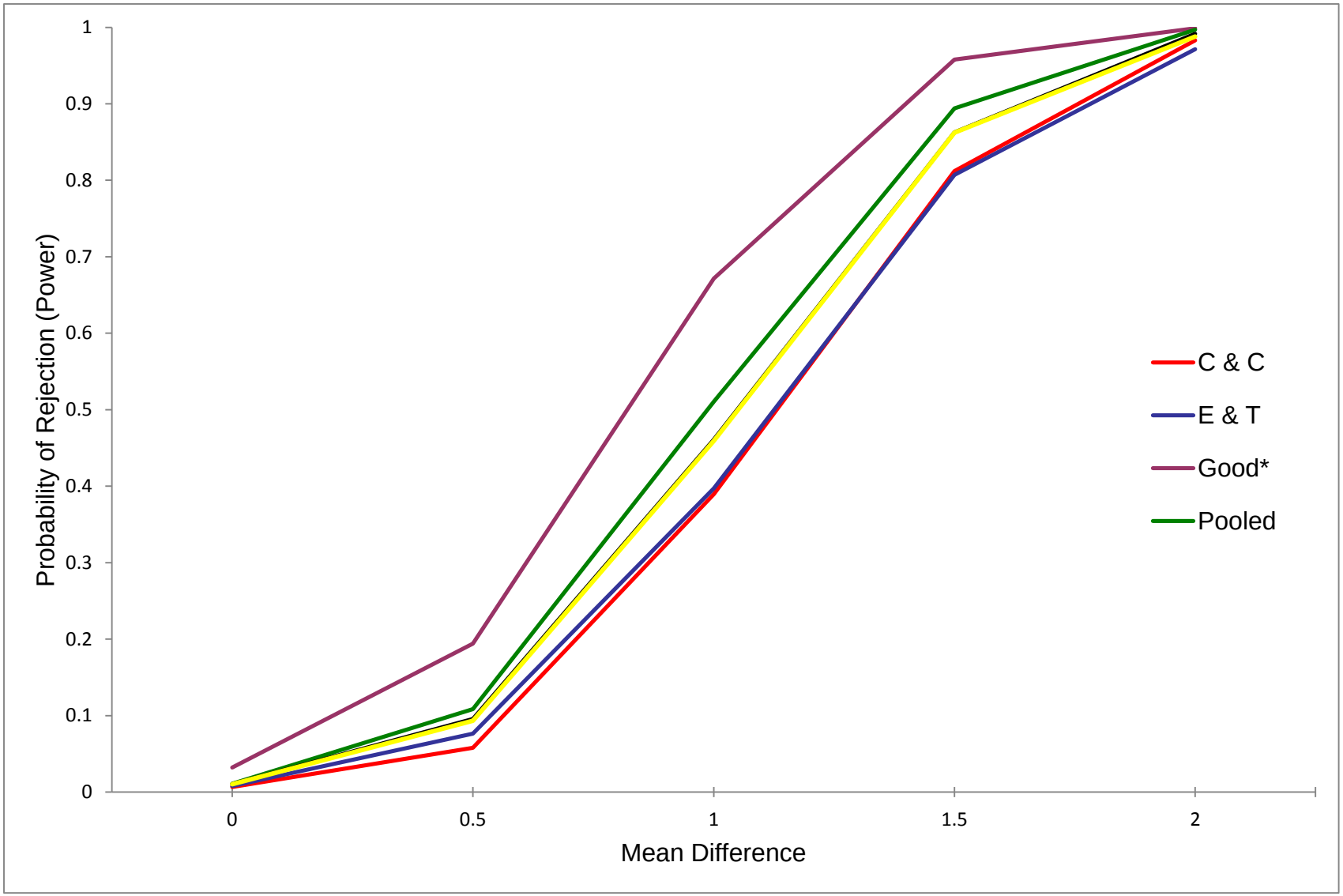

Figure A4. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

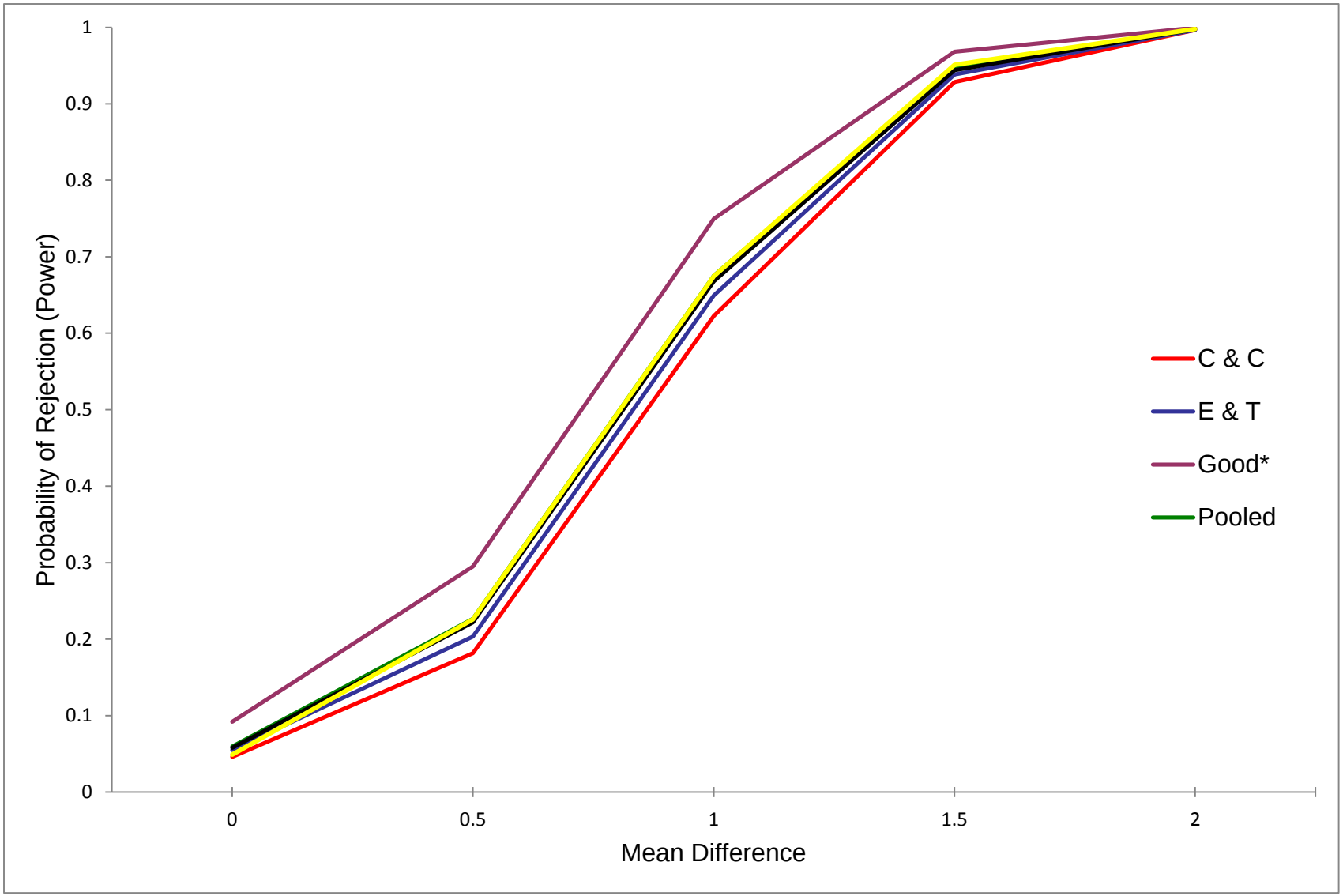

Figure A5. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

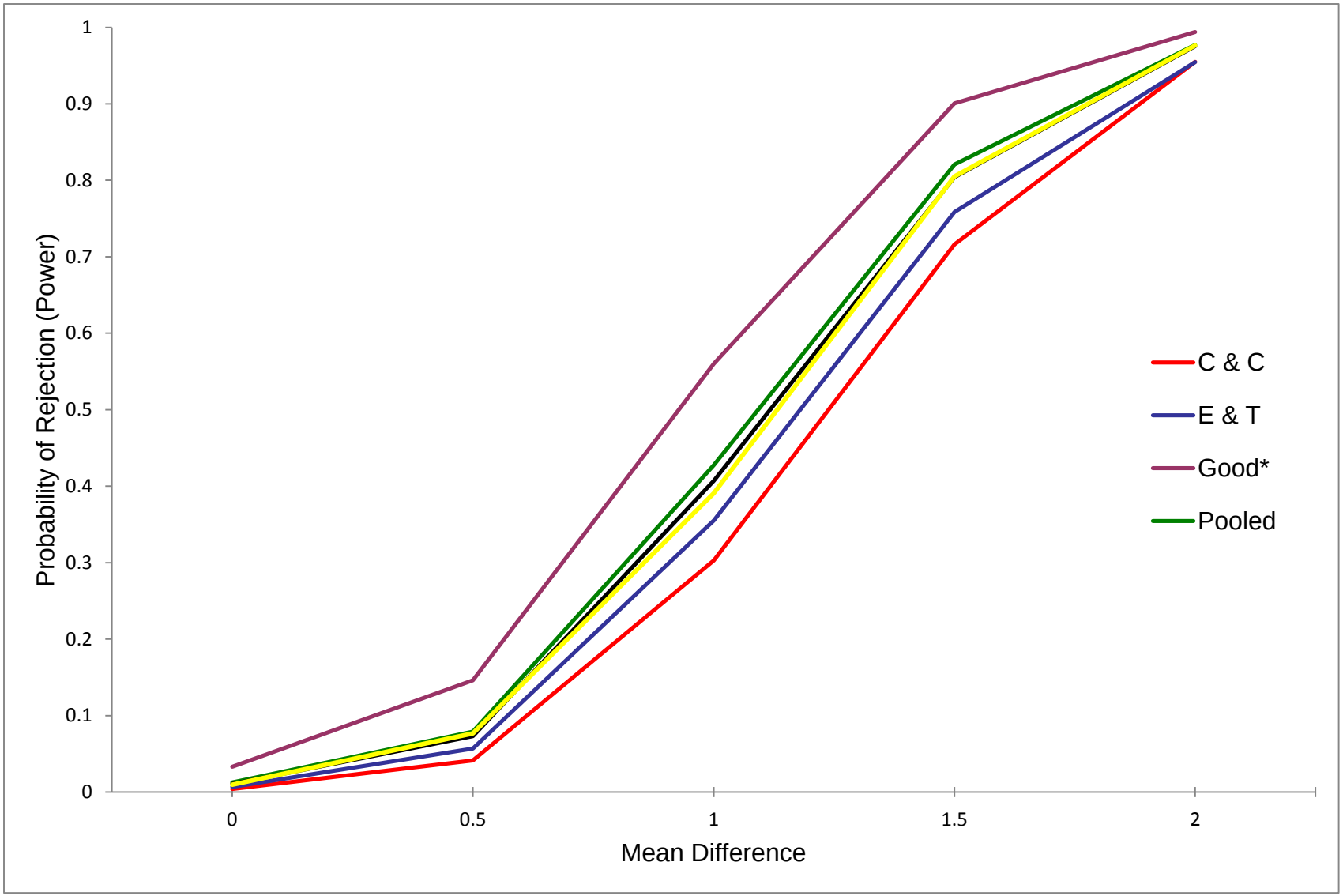

Figure A6. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

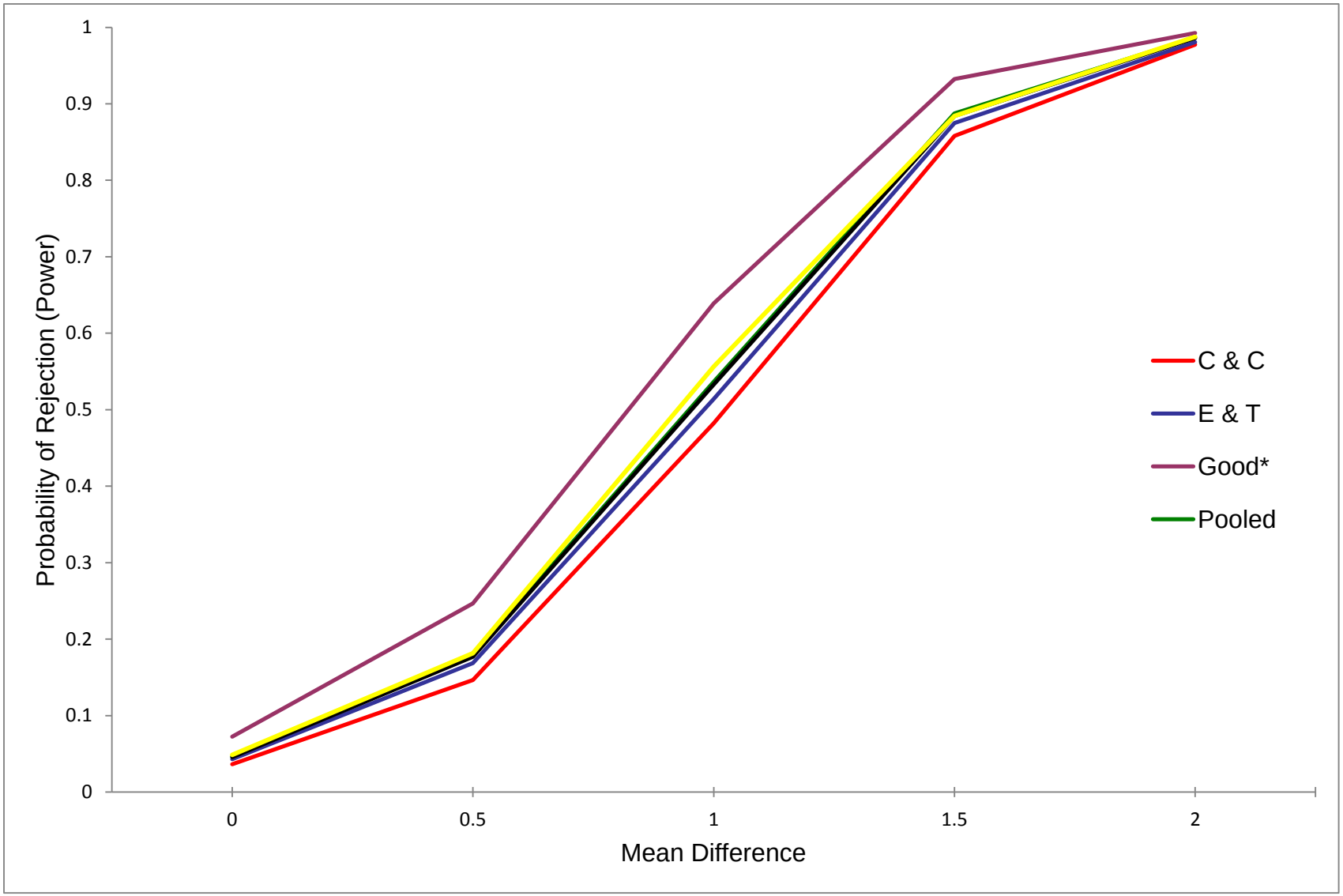

Figure A7. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

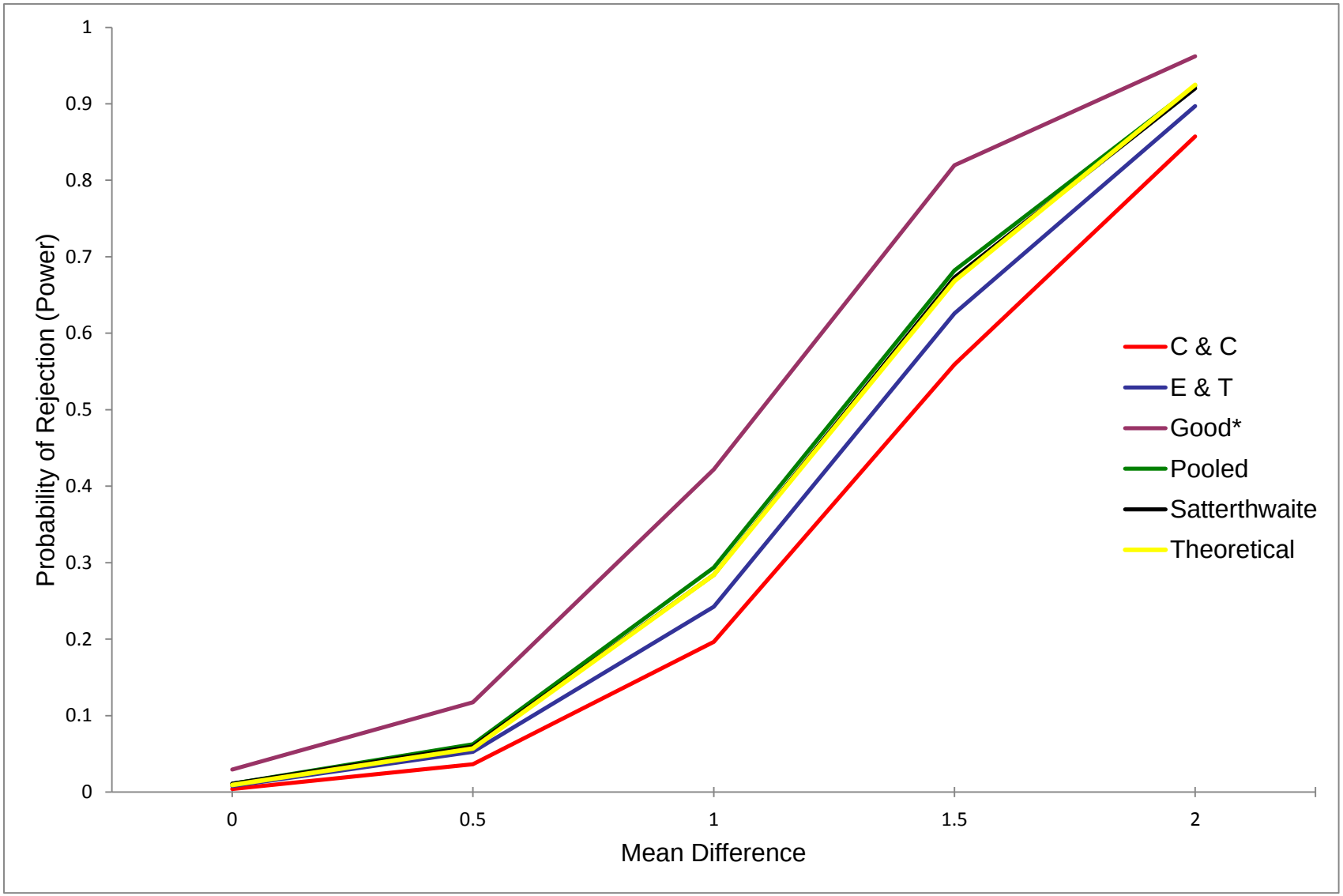

Figure A8. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

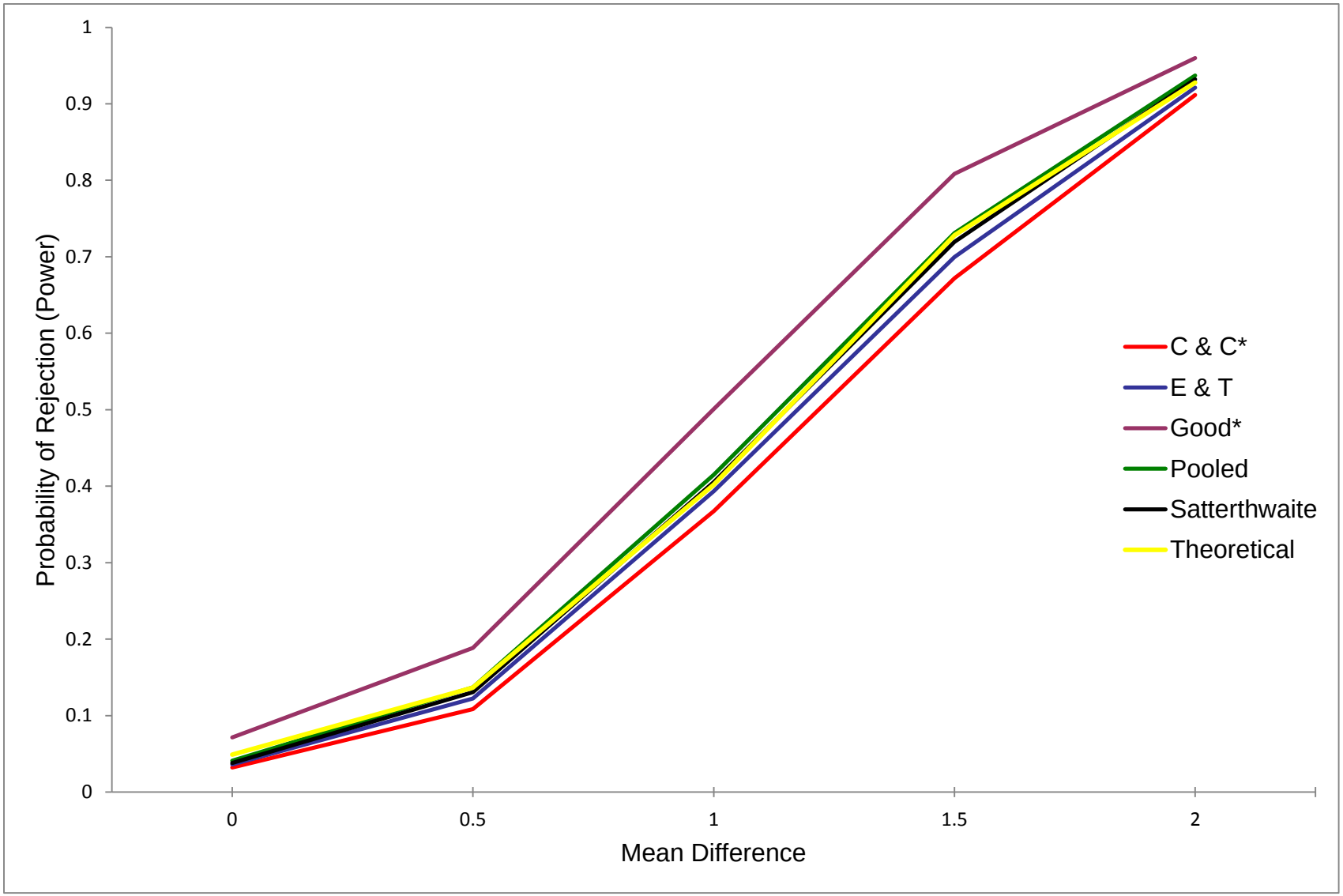

Figure A9. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

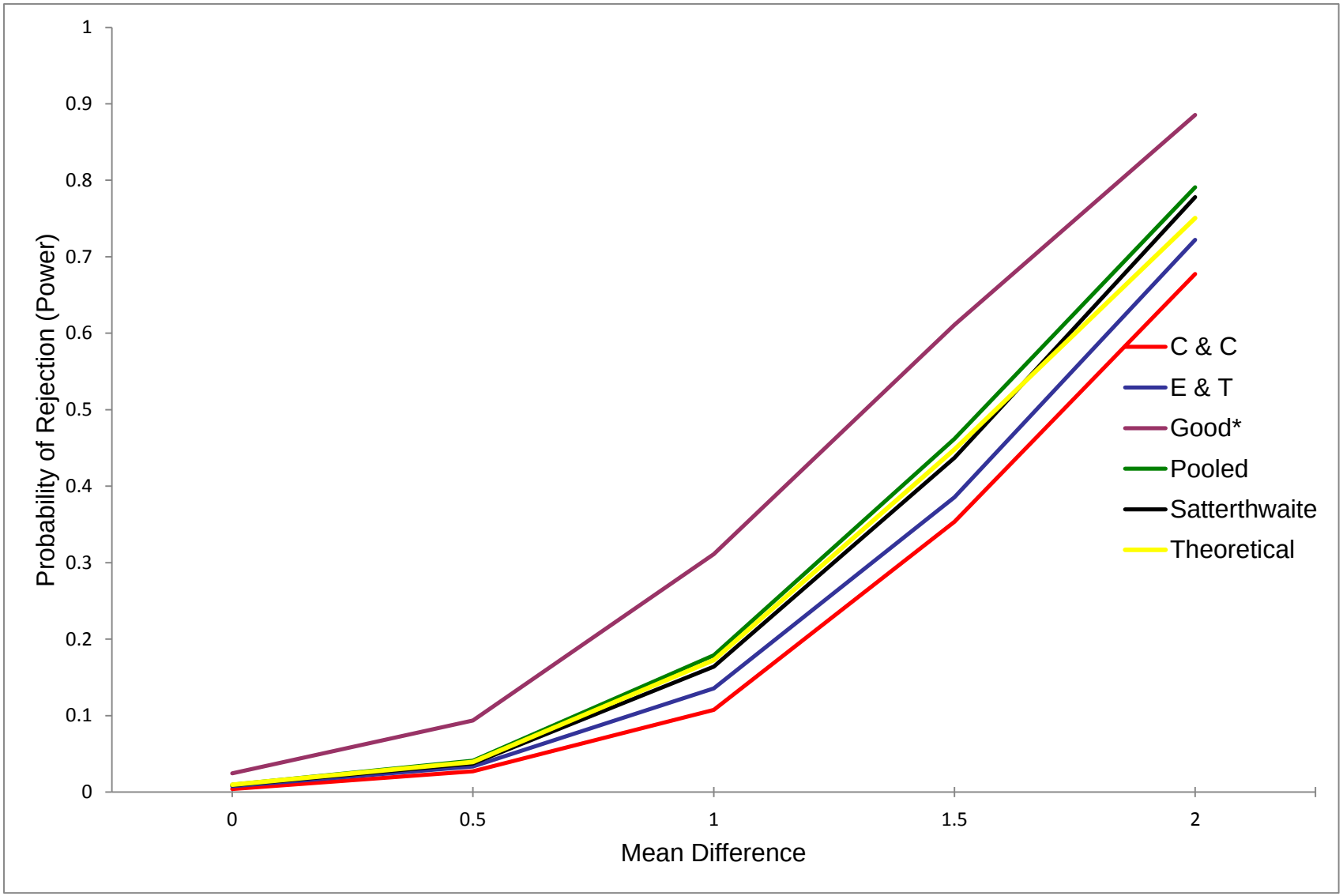

Figure A10 Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

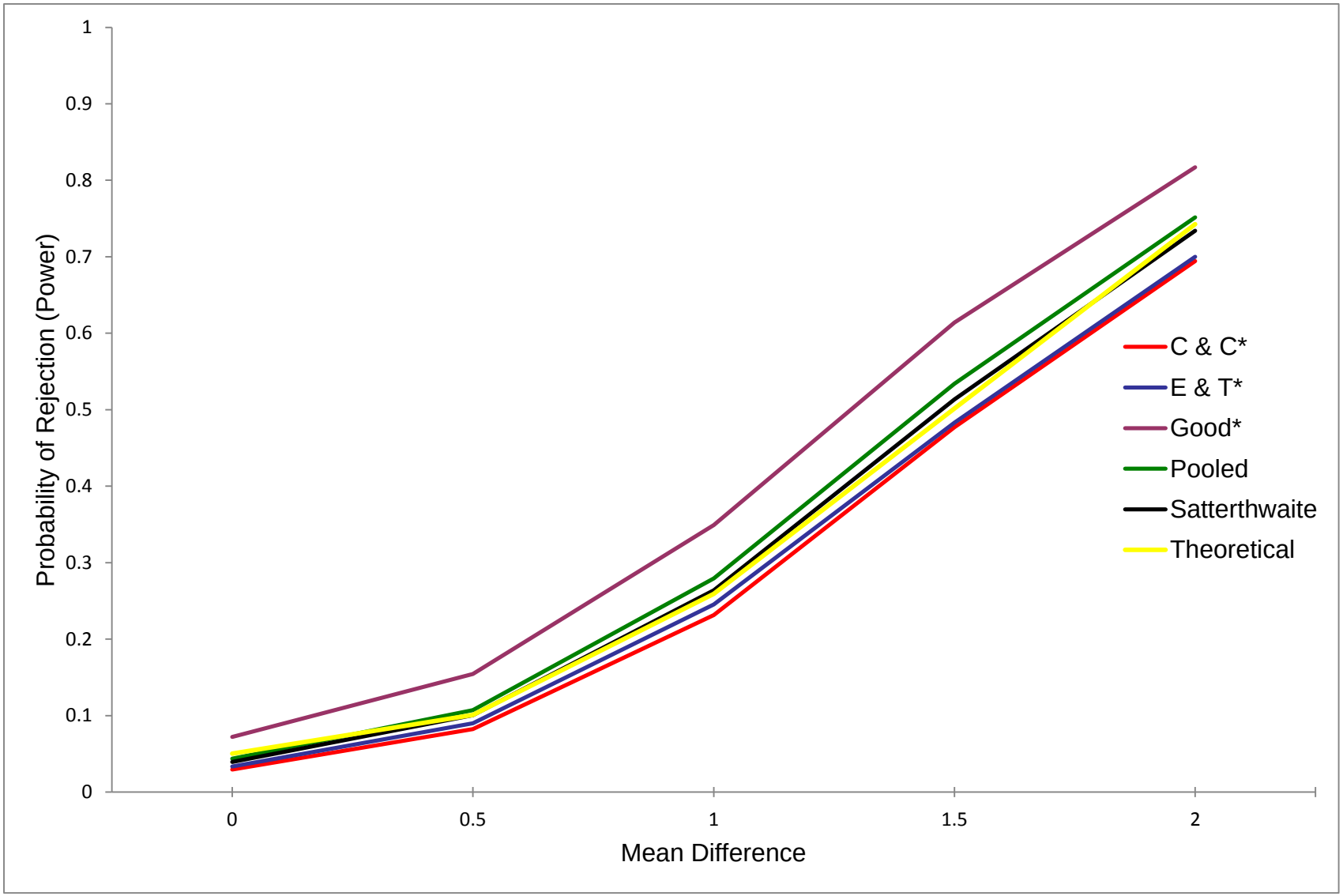

Figure A11. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

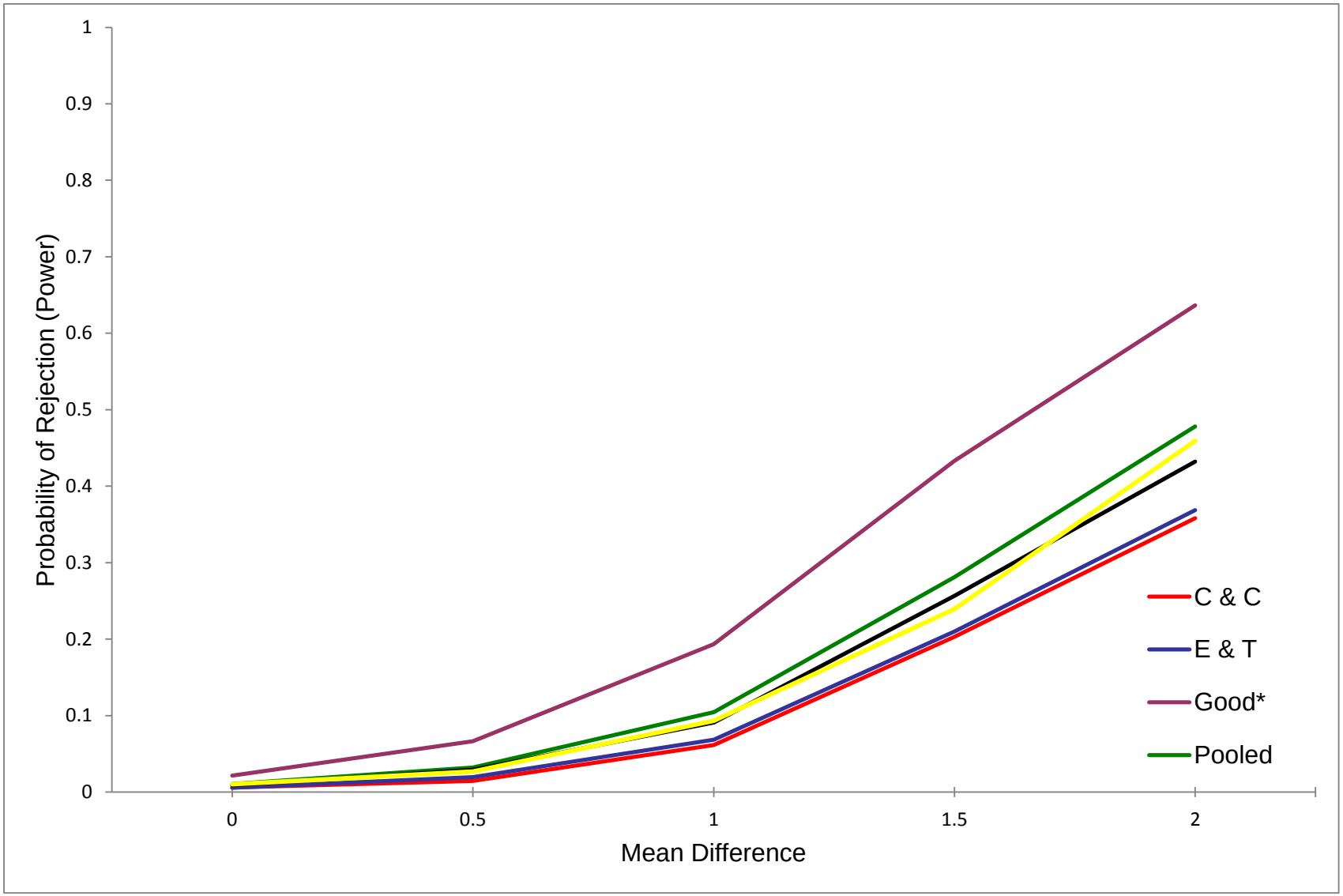

Figure A12. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

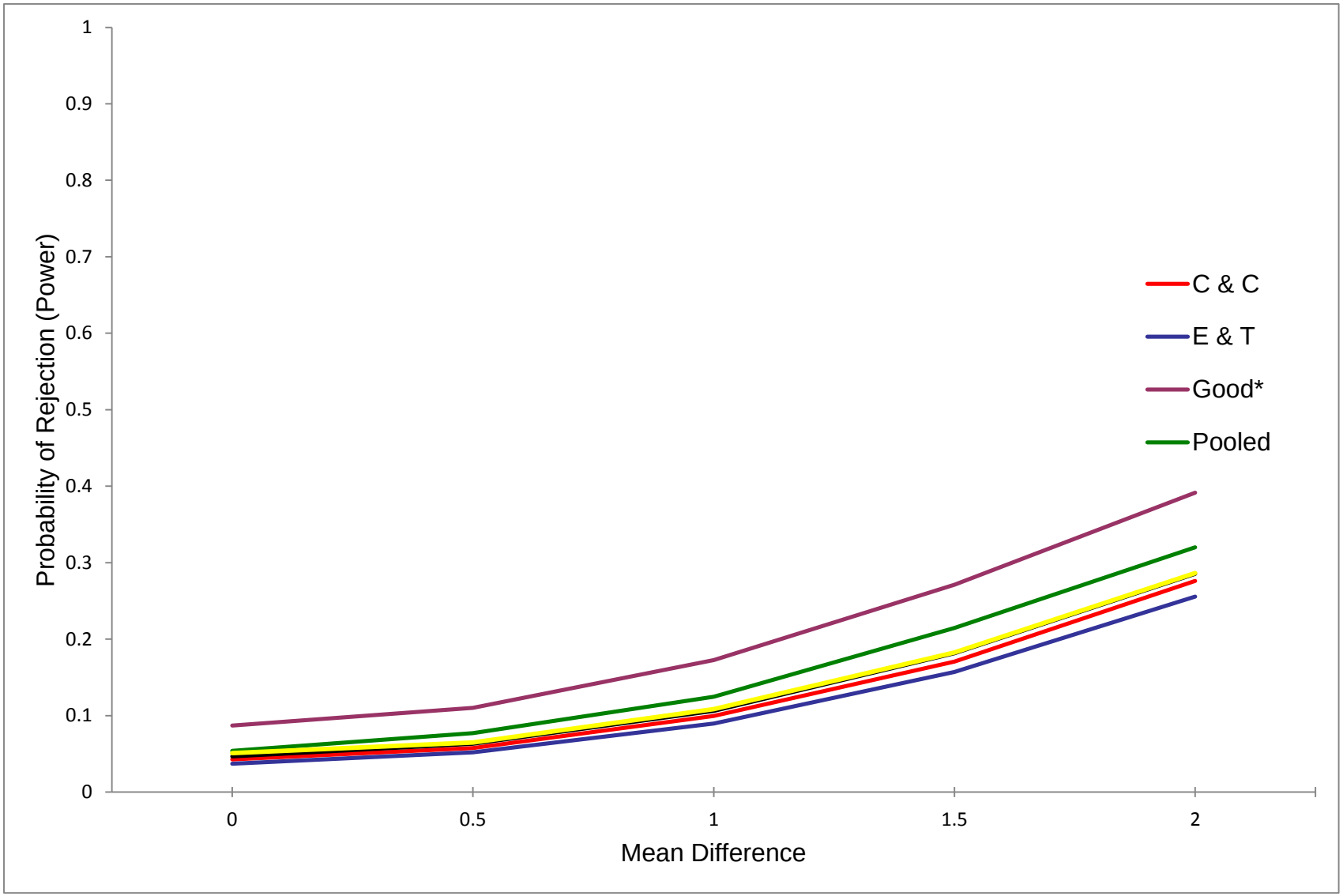

Figure A13. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

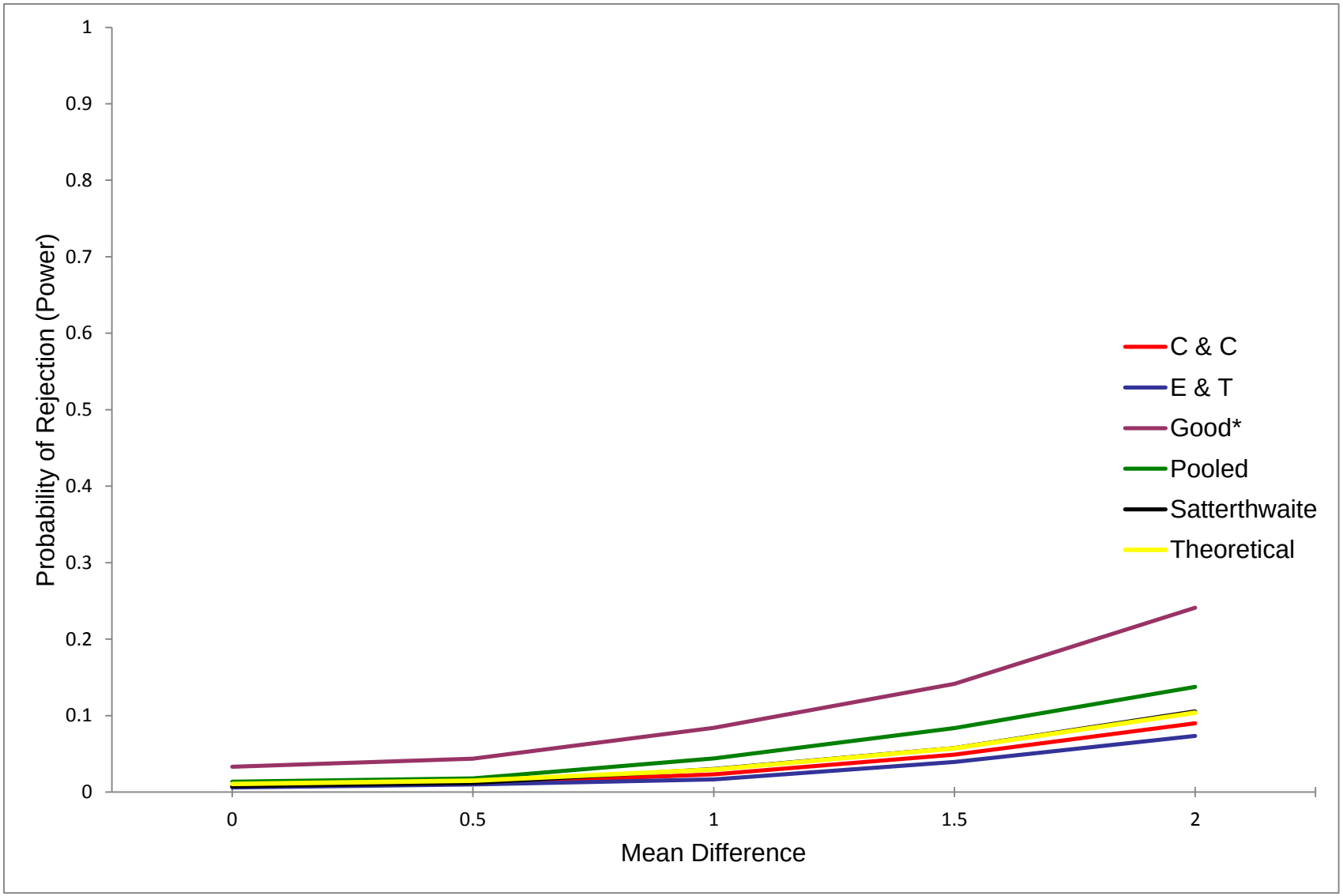

Figure A14. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 10$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 1.5 (i.e., $n_1 = 10$, $n_2 = 15$)

Table A10

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/16, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0590	0.0095
E & T	0.0600	0.0105
Good	0.0925*	0.0350*
Pooled	0.0355	0.0045
Satterthwaite	0.0610	0.0135

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A11

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0420	0.0060
E & T	0.0505	0.0090
Good	0.0745*	0.0225*
Pooled	0.0320*	0.0050
Satterthwaite	0.0530	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A12

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0450	0.0045
E & T	0.0545	0.0080
Good	0.0820*	0.0275*
Pooled	0.0460	0.0065
Satterthwaite	0.0565	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A13

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0380	0.0075
E & T	0.0435	0.0085
Good	0.0720*	0.0230*
Pooled	0.0465	0.0110
Satterthwaite	0.0440	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A14

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0395	0.0050
E & T	0.0440	0.0060
Good	0.0765*	0.0275*
Pooled	0.0640	0.0170
Satterthwaite	0.0470	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A15

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0365	0.0035
E & T	0.0355	0.0045
Good	0.0770*	0.0290*
Pooled	0.0760*	0.0205*
Satterthwaite	0.0410	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A16

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0505	0.0080
E & T	0.0425	0.0075
Good	0.0960*	0.0405*
Pooled	0.1140*	0.0395*
Satterthwaite	0.0525	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A17

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0590	0.0420	0.0450	0.0380	0.0395	0.0365	0.0505
	0.5	0.4030	0.2975	0.2515	0.1780	0.1315	0.0985	0.0785
	1.0	0.9270	0.8525	0.7605	0.5895	0.4015	0.2455	0.0995
	1.5	1.0000	0.9965	0.9830	0.9190	0.7410	0.5180	0.1805
	2.0	1.0000	1.0000	1.0000	0.9935	0.9485	0.7470	0.2745
Efron & Tibshirani	0.0	0.0600	0.0505	0.0545	0.0435	0.0440	0.0355	0.0425
	0.5	0.4000	0.3260	0.2805	0.2025	0.1405	0.1040	0.0665
	1.0	0.9260	0.8635	0.7910	0.6165	0.4170	0.2440	0.0910
	1.5	1.0000	0.9970	0.9860	0.9330	0.7525	0.5095	0.1630
	2.0	1.0000	1.0000	1.0000	0.9945	0.9480	0.7330	0.2445
Good	0.0	0.0925	0.0745	0.0820	0.0720	0.0765	0.0770	0.0960
	0.5	0.5010	0.4060	0.3510	0.2865	0.2130	0.1645	0.1325
	1.0	0.9615	0.9040	0.8480	0.7110	0.5405	0.3580	0.1885
	1.5	1.0000	0.9995	0.9935	0.9635	0.8475	0.6445	0.2795
	2.0	1.0000	1.0000	1.0000	0.9970	0.9800	0.8440	0.4125
Pooled	0.0	0.0355	0.0320	0.0460	0.0465	0.0640	0.0760	0.1140
	0.5	0.2945	0.2500	0.2490	0.2175	0.1895	0.1640	0.1515
	1.0	0.8725	0.8135	0.7585	0.6470	0.5110	0.3740	0.2150
	1.5	0.9980	0.9940	0.9850	0.9420	0.8315	0.6650	0.3155
	2.0	1.0000	1.0000	1.0000	0.9965	0.9770	0.8615	0.4640
Satterthwaite	0.0	0.0610	0.0530	0.0565	0.0440	0.0470	0.0410	0.0525
	0.5	0.4165	0.3360	0.2905	0.2125	0.1500	0.1115	0.0800
	1.0	0.9310	0.8695	0.8010	0.6315	0.4400	0.2625	0.1050
	1.5	1.0000	0.9975	0.9880	0.9385	0.7705	0.5375	0.1870
	2.0	1.0000	1.0000	1.0000	0.9950	0.9560	0.7670	0.2820

Table A18

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 15$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0095	0.0060	0.0045	0.0075	0.0050	0.0035	0.0080
	0.5	0.1570	0.1015	0.0720	0.0455	0.0310	0.0220	0.0170
	1.0	0.7480	0.5885	0.4340	0.2890	0.1420	0.0815	0.0250
	1.5	0.9910	0.9660	0.8805	0.7015	0.4180	0.2145	0.0480
	2.0	1.0000	1.0000	0.9920	0.9420	0.7595	0.4200	0.1030
Efron & Tibshirani	0.0	0.0105	0.0090	0.0080	0.0085	0.0060	0.0045	0.0075
	0.5	0.1535	0.1185	0.0935	0.0570	0.0390	0.0225	0.0125
	1.0	0.7235	0.6275	0.5010	0.3350	0.1620	0.0820	0.0220
	1.5	0.9845	0.9730	0.9075	0.7440	0.4590	0.2145	0.0440
	2.0	0.9990	0.9995	0.9965	0.9490	0.7635	0.4090	0.0775
Good	0.0	0.0350	0.0225	0.0275	0.0230	0.0275	0.0290	0.0405
	0.5	0.2995	0.2150	0.1800	0.1365	0.0915	0.0755	0.0615
	1.0	0.8740	0.7750	0.6685	0.5085	0.3340	0.2060	0.0815
	1.5	0.9975	0.9915	0.9660	0.8800	0.6800	0.4465	0.1455
	2.0	1.0000	1.0000	0.9990	0.9840	0.9215	0.6895	0.2345
Pooled	0.0	0.0045	0.0050	0.0065	0.0110	0.0170	0.0205	0.0395
	0.5	0.0915	0.0875	0.0835	0.0745	0.0675	0.0615	0.0660
	1.0	0.6285	0.5525	0.4700	0.3835	0.2755	0.1865	0.0870
	1.5	0.9740	0.9515	0.8995	0.8040	0.6150	0.4215	0.1615
	2.0	0.9990	0.9995	0.9935	0.9690	0.9000	0.6735	0.2425
Satterthwaite	0.0	0.0135	0.0110	0.0110	0.0100	0.0075	0.0075	0.0105
	0.5	0.1850	0.1330	0.1080	0.0715	0.0465	0.0260	0.0185
	1.0	0.7665	0.6615	0.5400	0.3705	0.1925	0.1005	0.0280
	1.5	0.9910	0.9790	0.9265	0.7815	0.4970	0.2560	0.0565
	2.0	1.0000	1.0000	0.9975	0.9640	0.8200	0.4730	0.1105

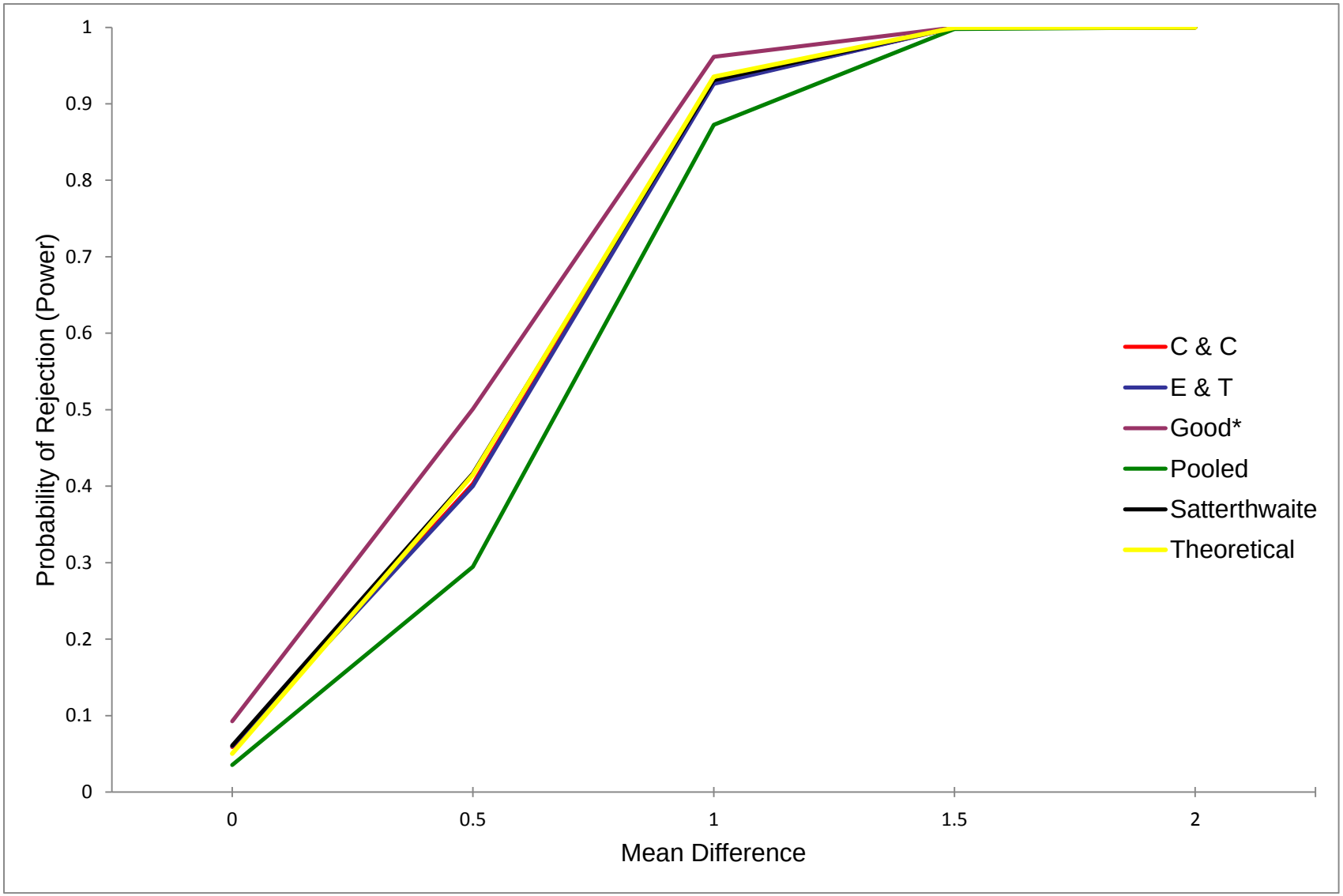

Figure A15. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

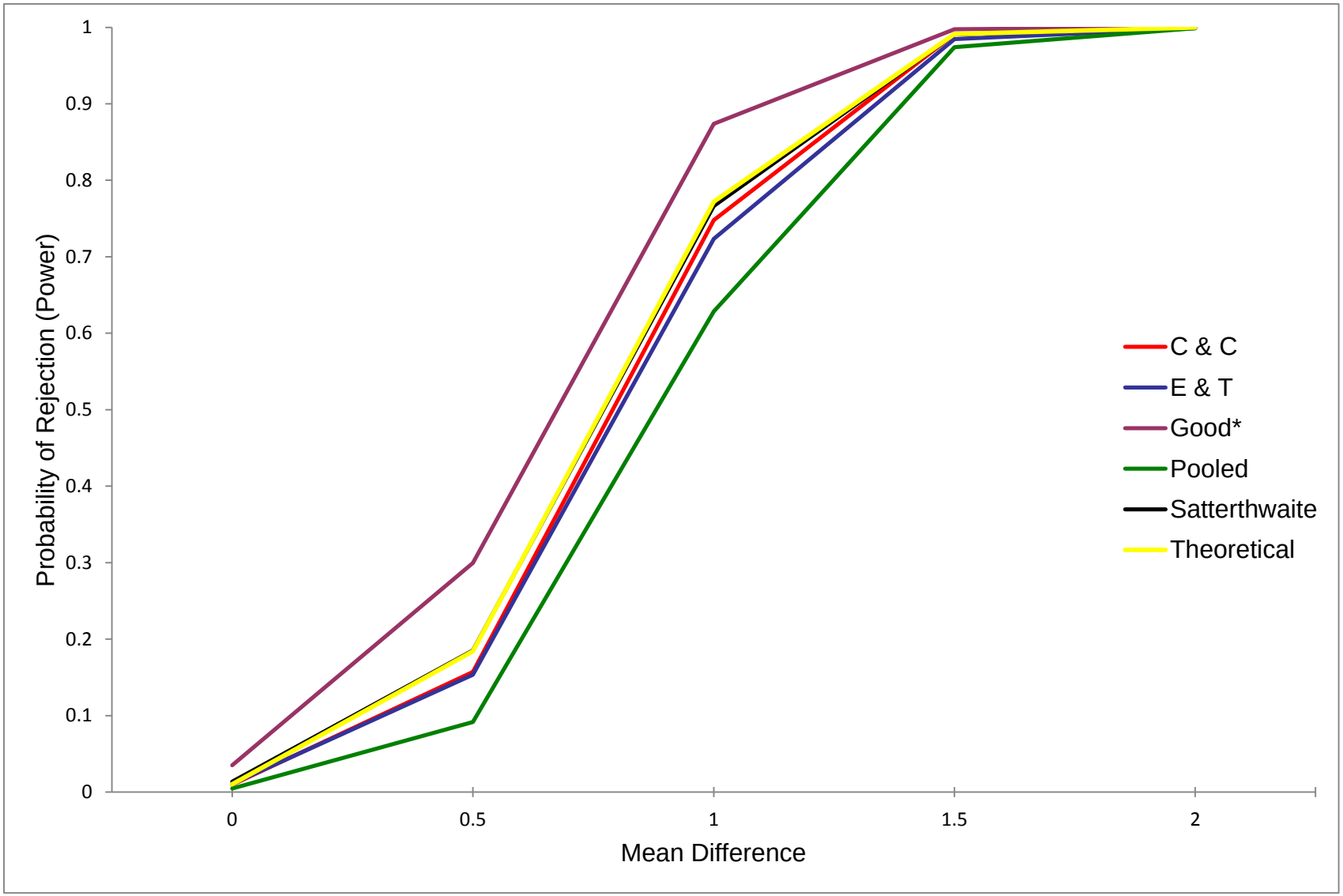

Figure A16. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

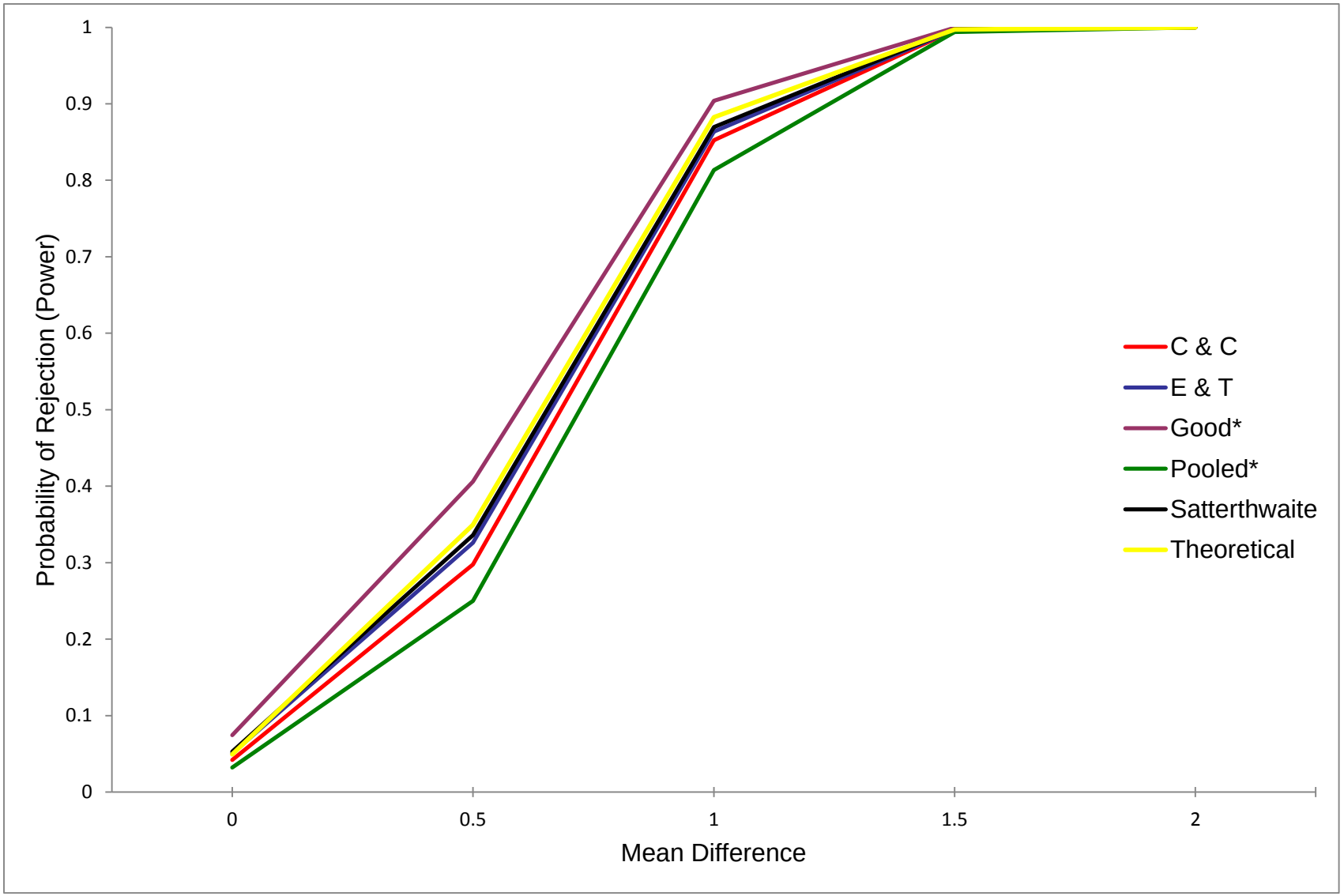

Figure A17. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

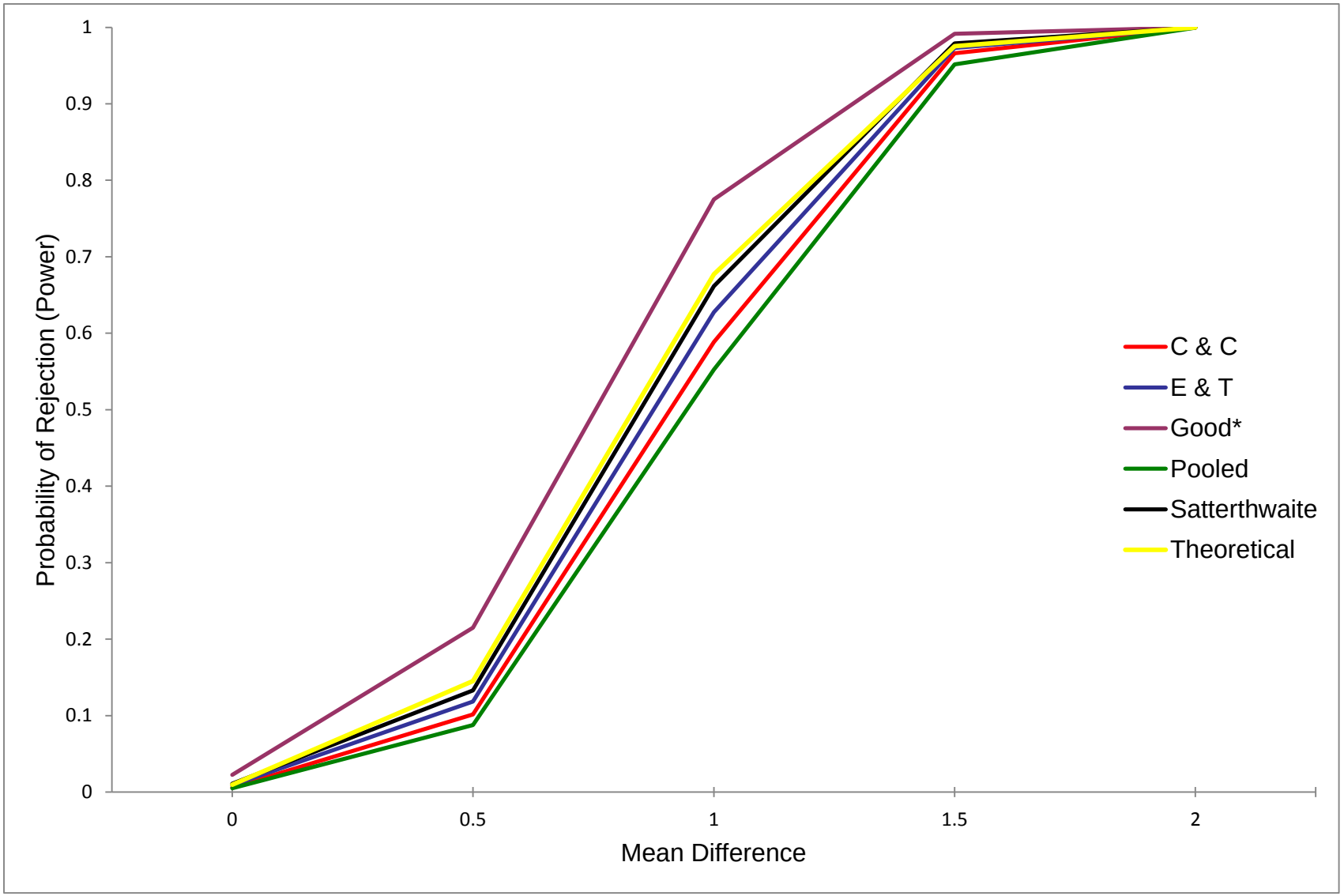

Figure A18. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

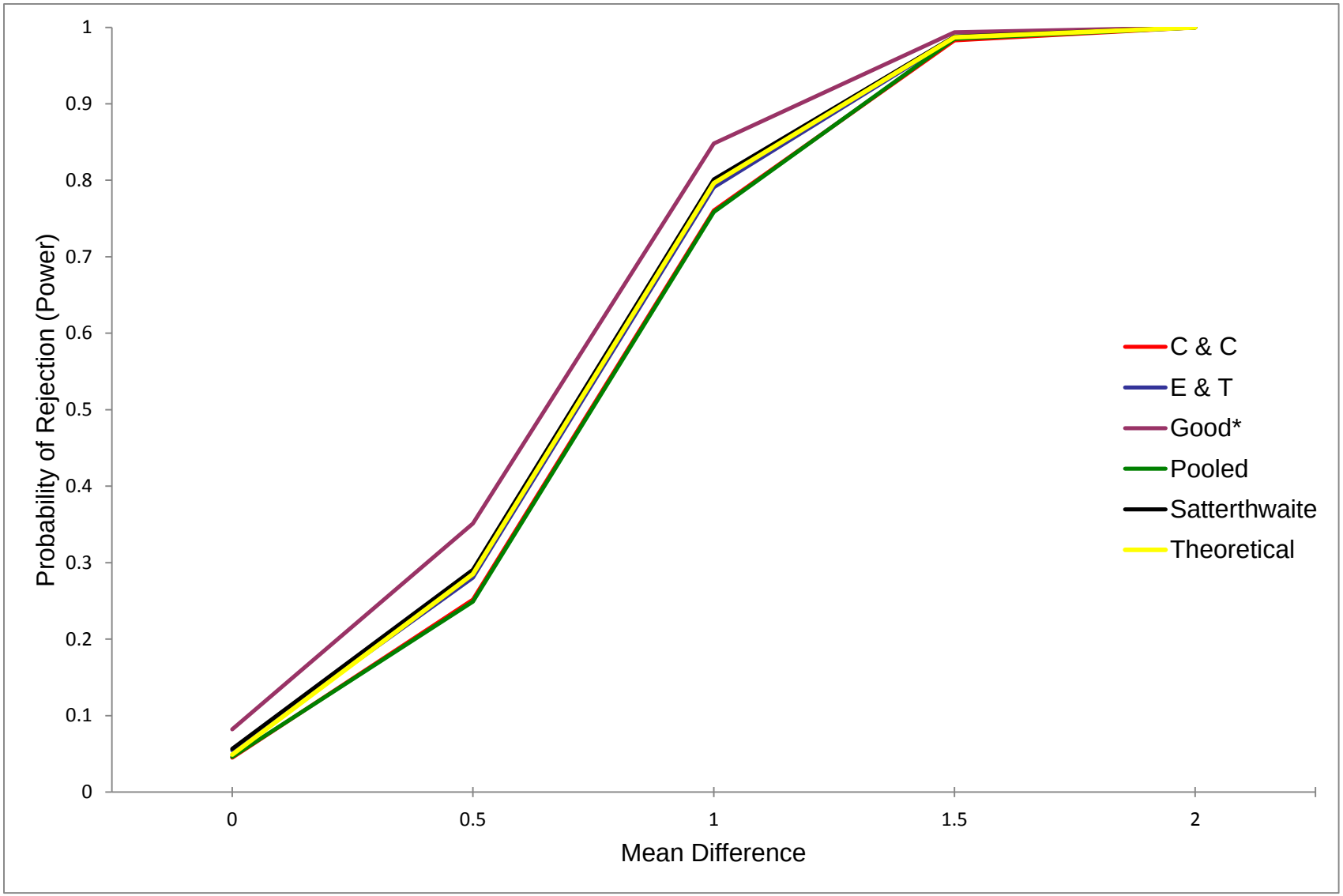

Figure A19. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

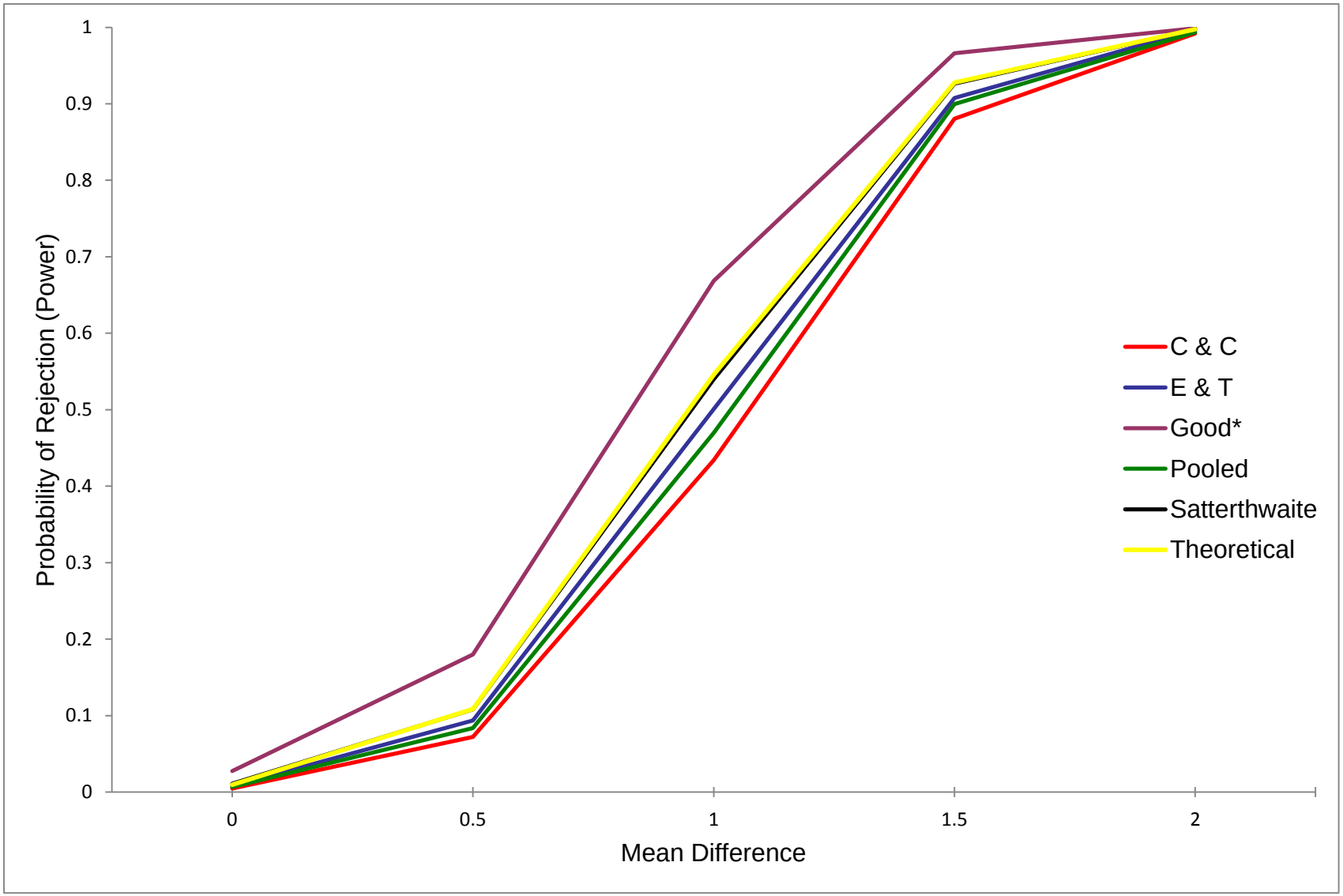

Figure A20. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

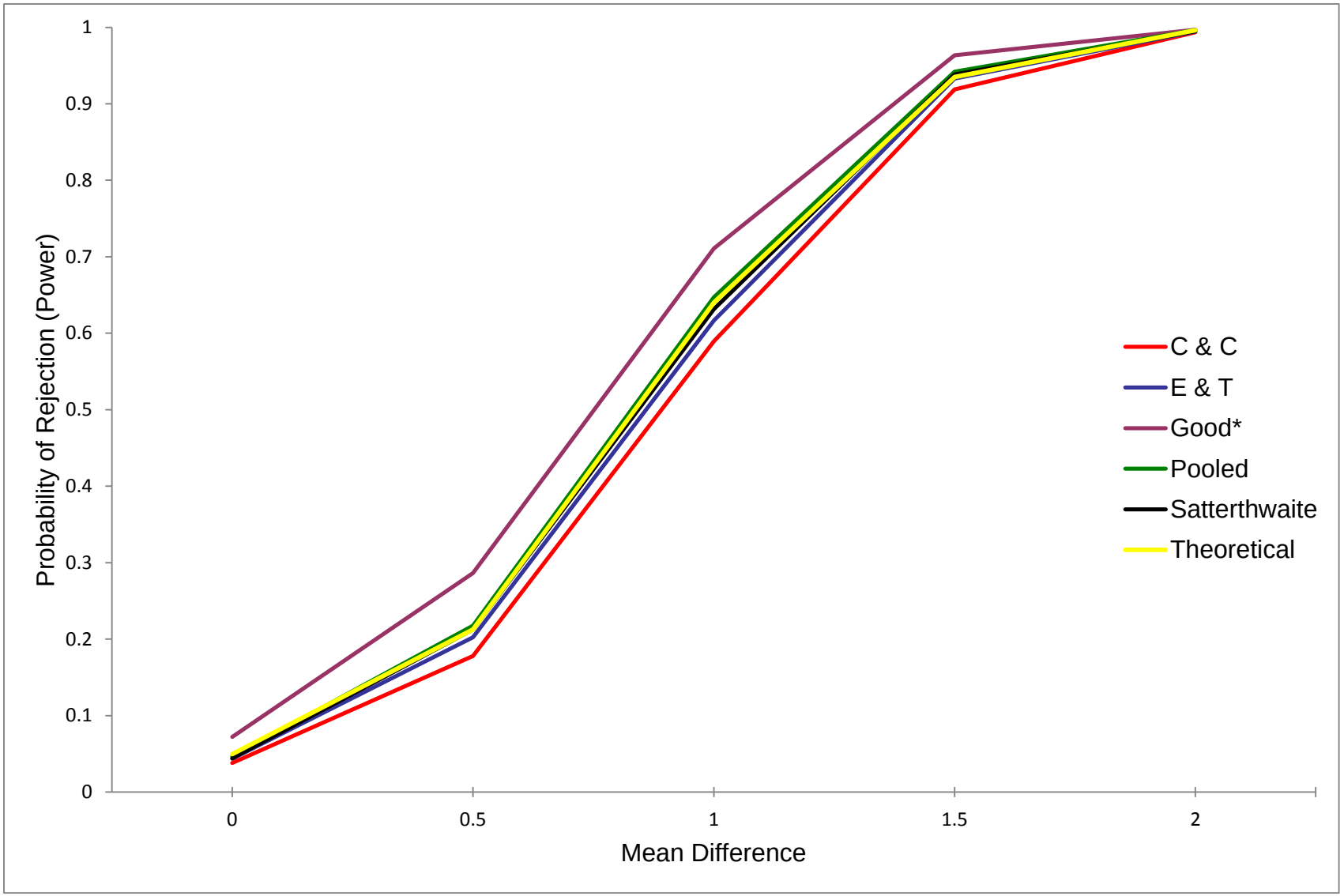

Figure A21. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

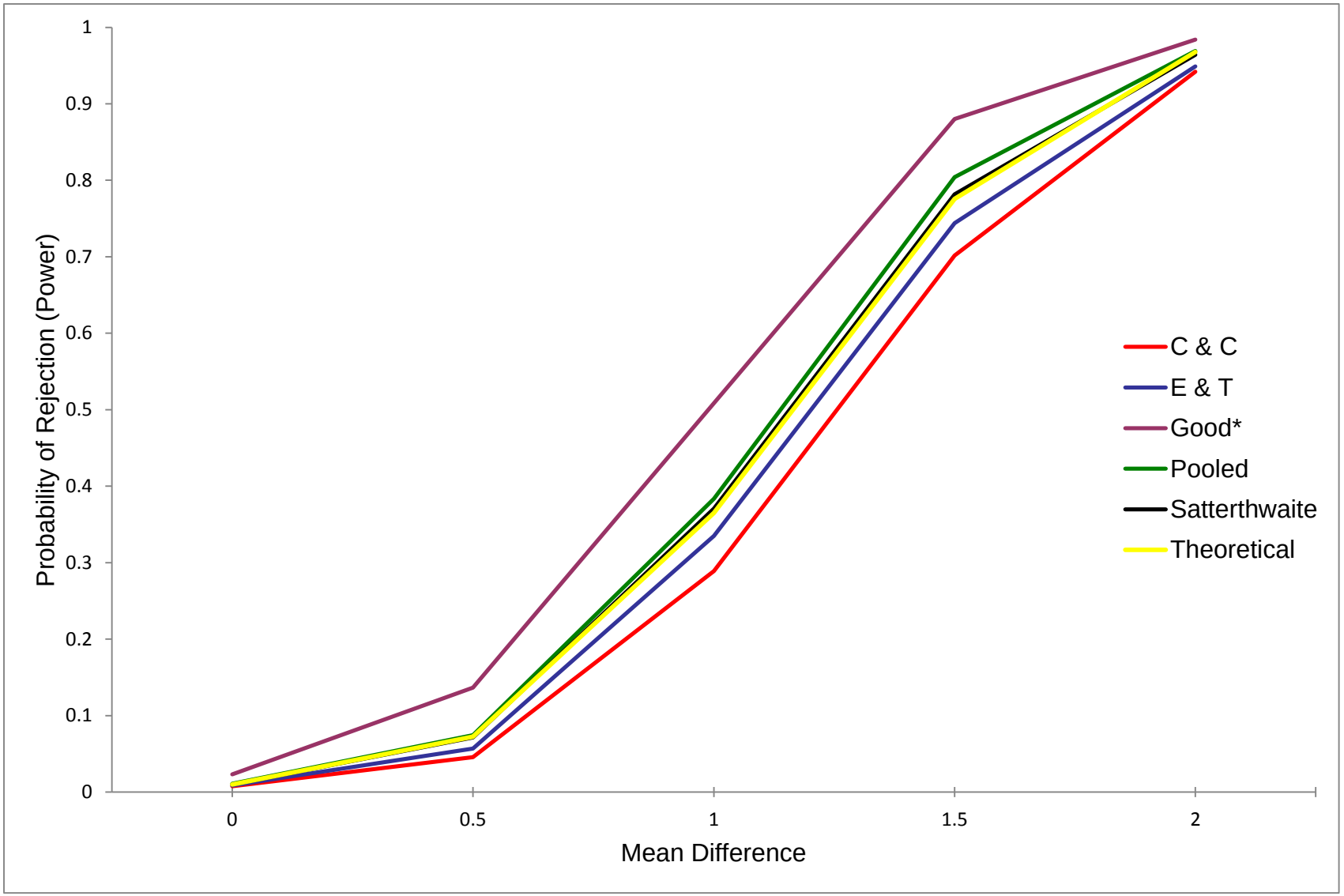

Figure A22. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

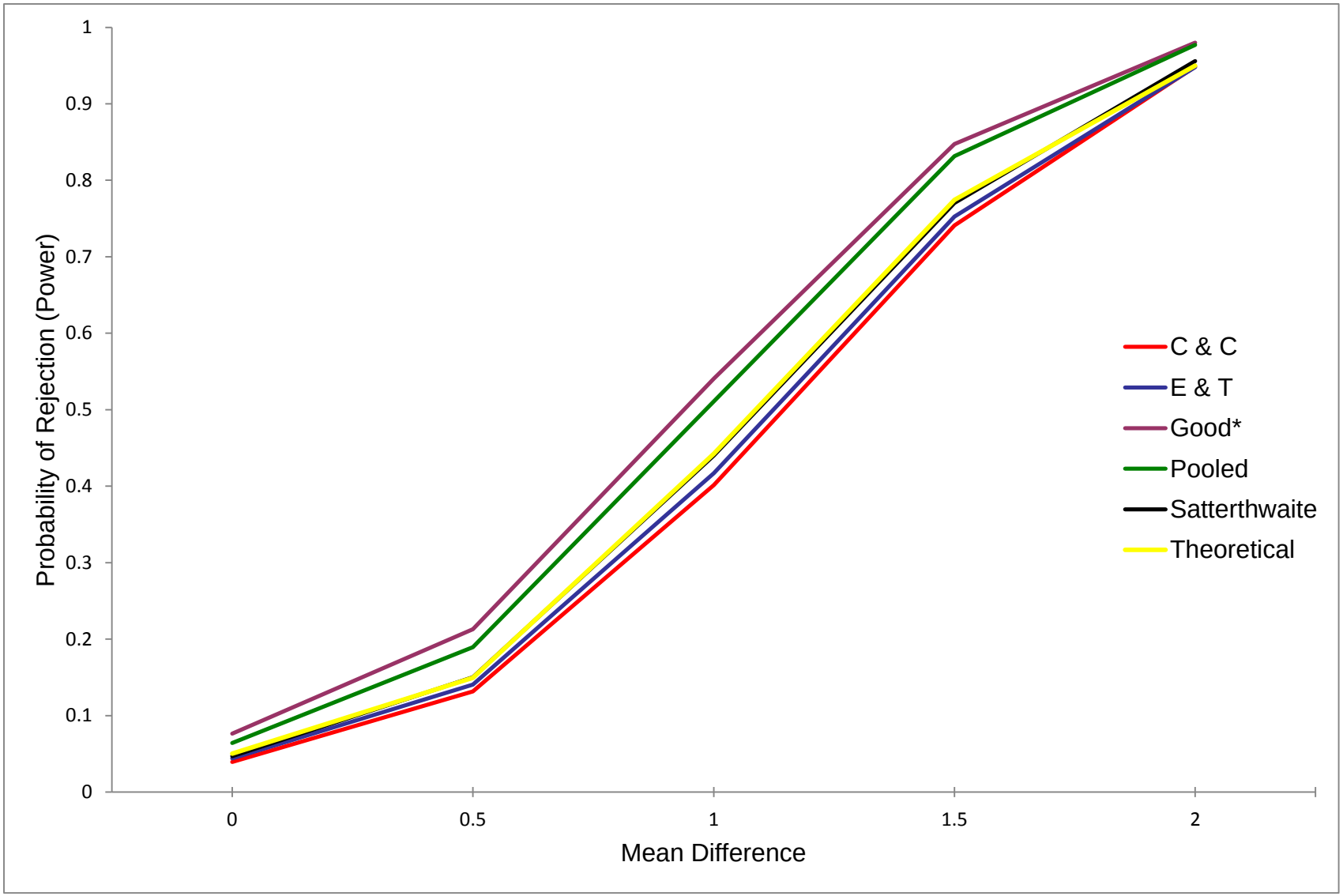

Figure A23. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

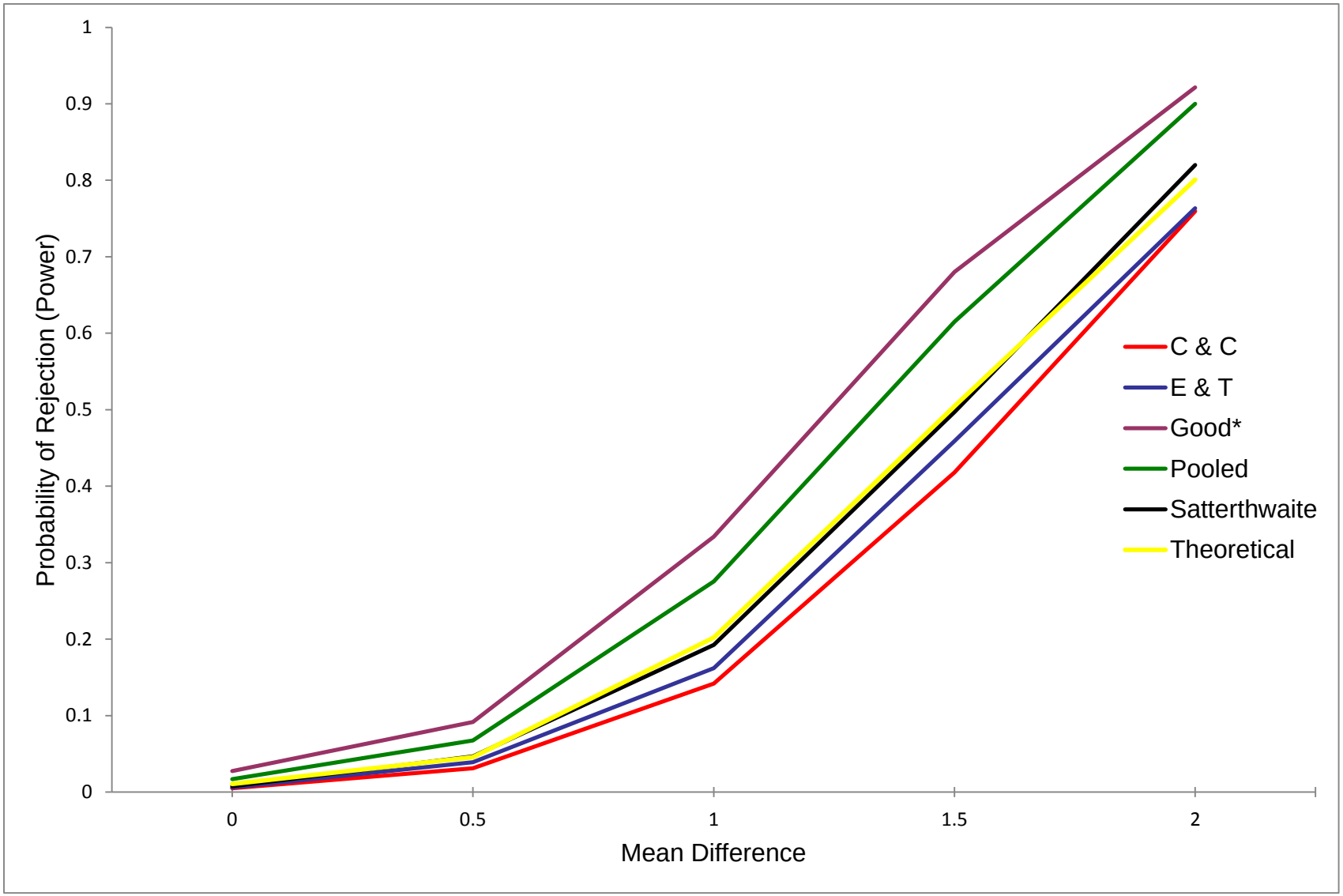

Figure A24. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

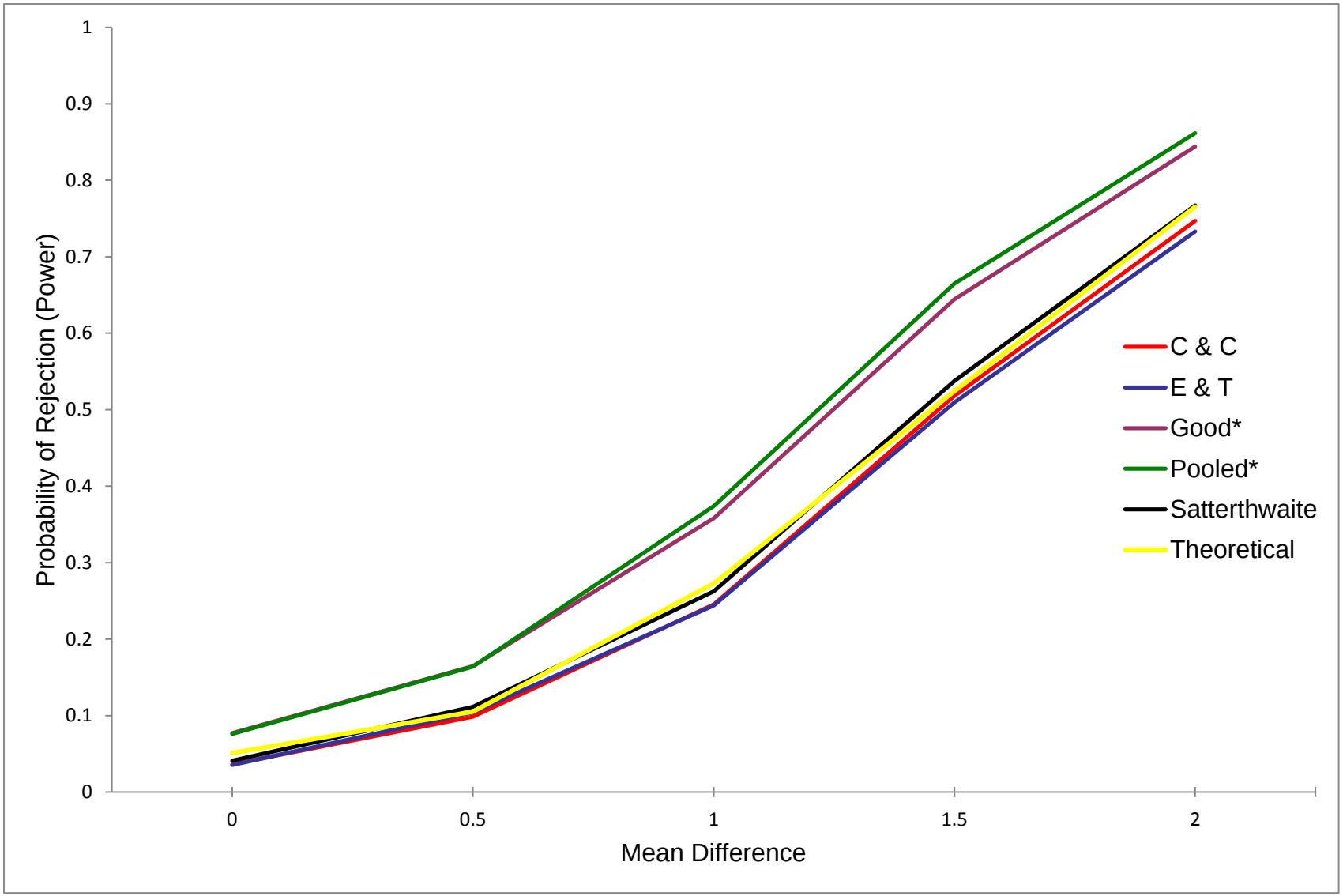

Figure A25. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

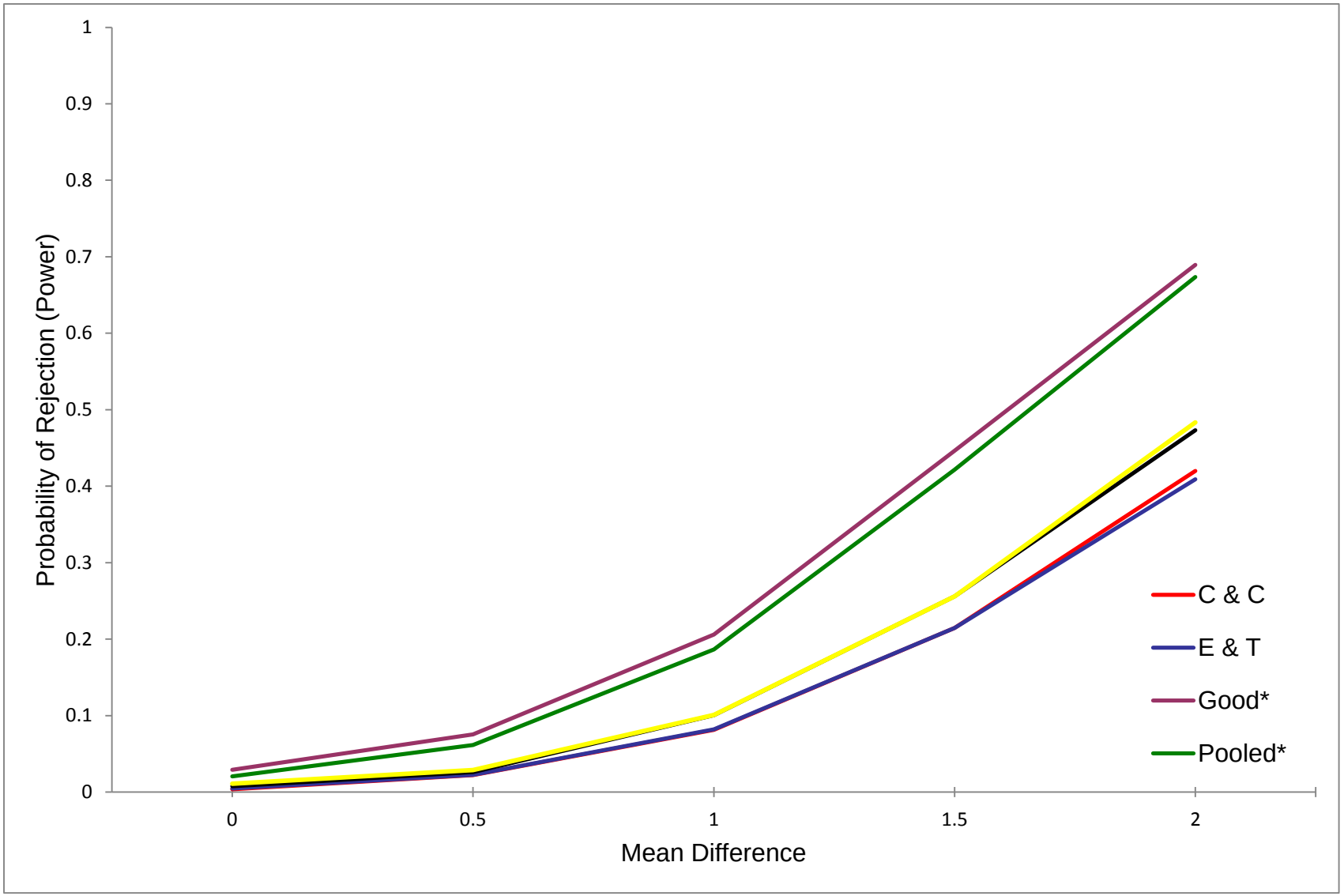

Figure A26. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

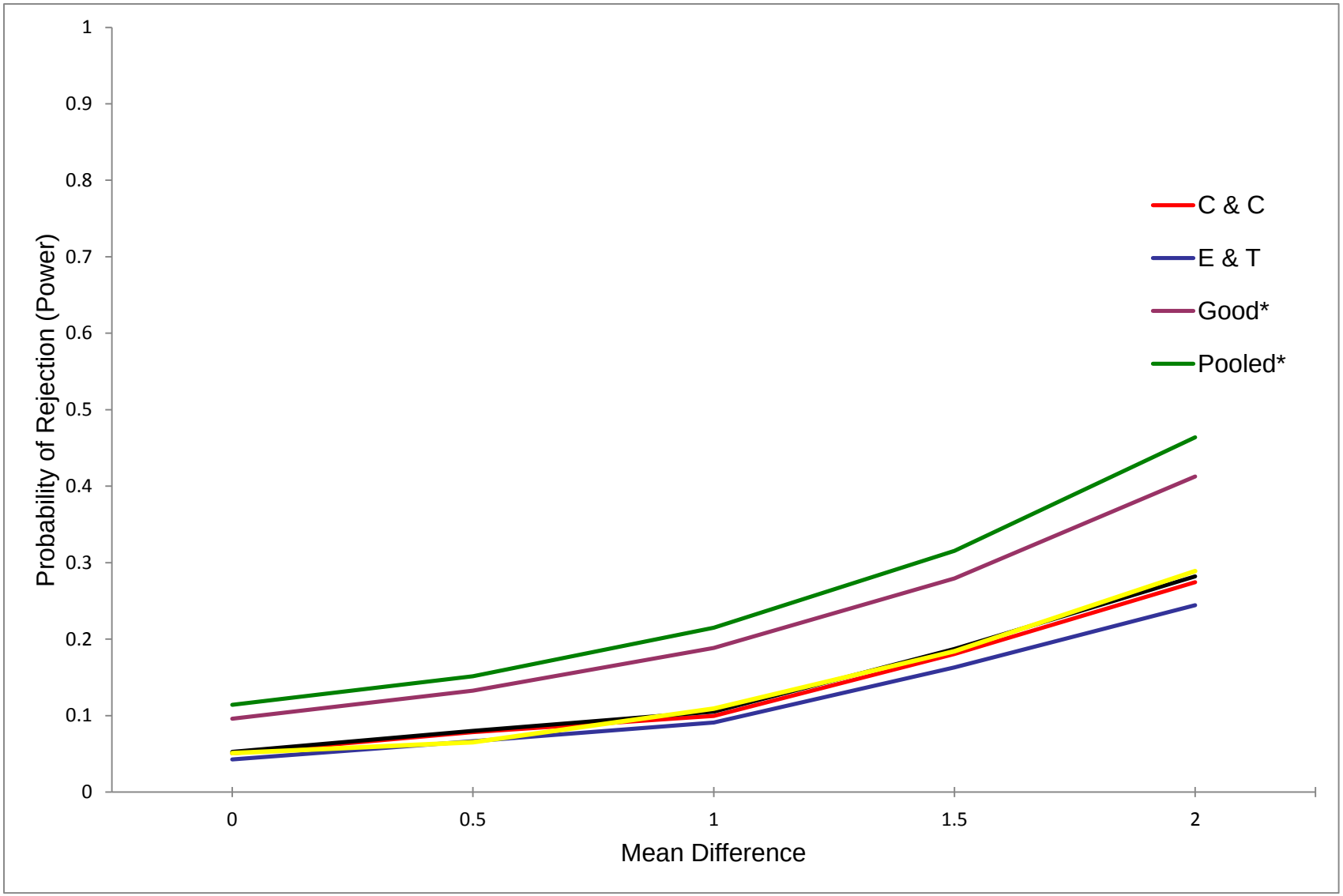

Figure A27. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

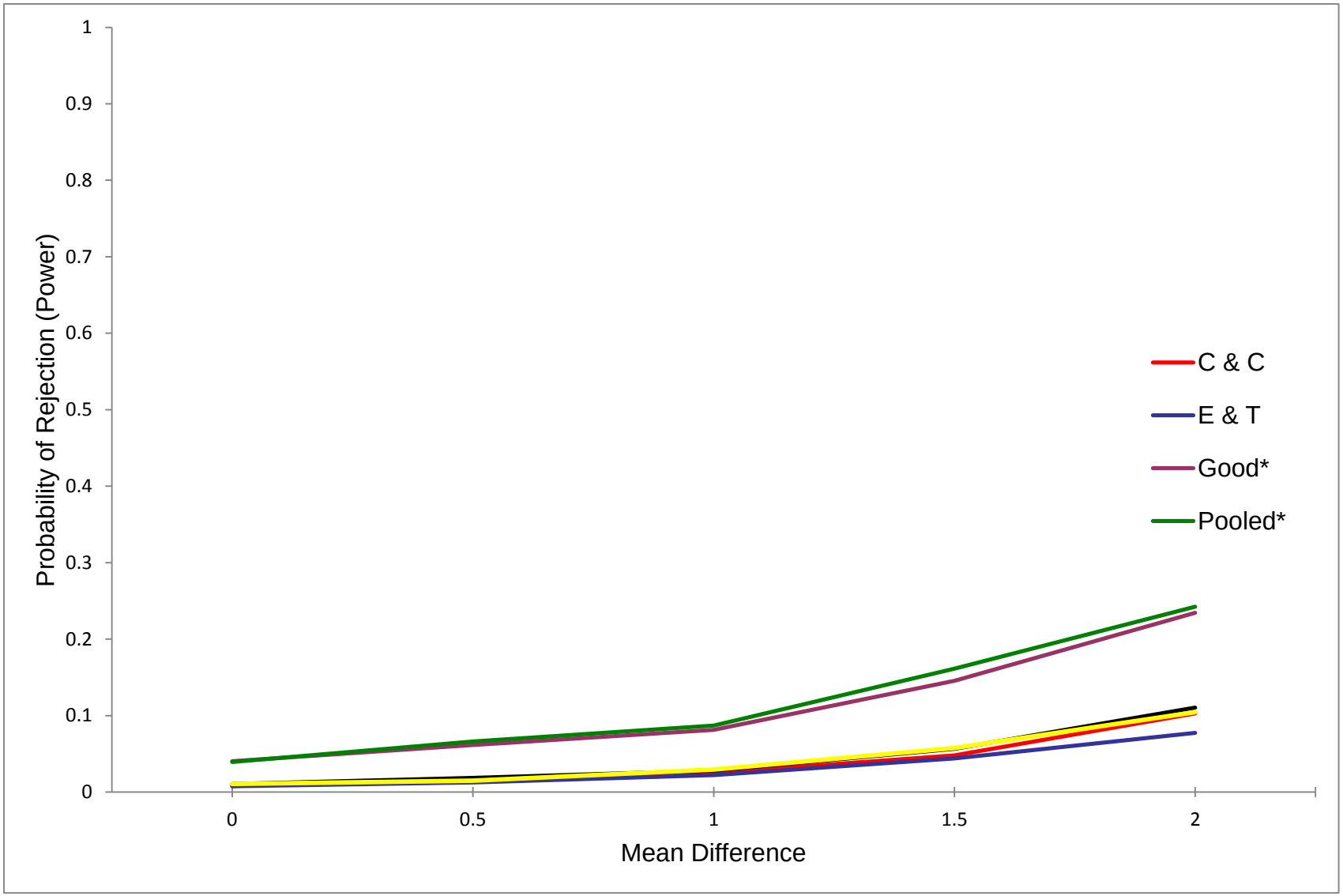

Figure A28. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 15$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 3.0 (i.e., $n_1 = 10$, $n_2 = 30$)

Table A19

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0400	0.0050
E & T	0.0430	0.0065
Good	0.0590	0.0150
Pooled	0.0015*	<.0005*
Satterthwaite	0.0445	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A20

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0405	0.0045
E & T	0.0525	0.0080
Good	0.0670*	0.0140
Pooled	0.0070*	0.0010*
Satterthwaite	0.0520	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A21

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0365	0.0045
E & T	0.0445	0.0055
Good	0.0640	0.0160
Pooled	0.0155*	0.0015*
Satterthwaite	0.0450	0.0065

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A22

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0375	0.0065
E & T	0.0445	0.0075
Good	0.0780*	0.0200*
Pooled	0.0445	0.0100
Satterthwaite	0.0465	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A23

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0420	0.0090
E & T	0.0425	0.0115
Good	0.0820*	0.0285*
Pooled	0.0990*	0.0275*
Satterthwaite	0.0465	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A24

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0465	0.0075
E & T	0.0440	0.0070
Good	0.0880*	0.0365*
Pooled	0.1535*	0.0620*
Satterthwaite	0.0490	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A25

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0415	0.0075
E & T	0.0295*	0.0050
Good	0.0845*	0.0350*
Pooled	0.2230*	0.1055*
Satterthwaite	0.0415	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A26

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0400	0.0405	0.0365	0.0375	0.0420	0.0465	0.0415
	0.5	0.6835	0.4615	0.3475	0.2360	0.1425	0.1065	0.0650
	1.0	0.9980	0.9670	0.9025	0.6990	0.4770	0.2775	0.1135
	1.5	1.0000	1.0000	0.9975	0.9705	0.8140	0.5260	0.1680
	2.0	1.0000	1.0000	1.0000	0.9995	0.9610	0.7675	0.2810
Efron & Tibshirani	0.0	0.0430	0.0525	0.0445	0.0445	0.0425	0.0440	0.0295
	0.5	0.6965	0.4900	0.3715	0.2505	0.1365	0.1000	0.0535
	1.0	0.9985	0.9710	0.9135	0.7050	0.4655	0.2570	0.0970
	1.5	1.0000	1.0000	0.9985	0.9705	0.8060	0.5000	0.1445
	2.0	1.0000	1.0000	1.0000	0.9980	0.9585	0.7380	0.2500
Good	0.0	0.0590	0.0670	0.0640	0.0780	0.0820	0.0880	0.0845
	0.5	0.7445	0.5605	0.4515	0.3380	0.2265	0.1765	0.1190
	1.0	0.9990	0.9810	0.9450	0.7995	0.6090	0.3905	0.1835
	1.5	1.0000	1.0000	0.9990	0.9865	0.8920	0.6635	0.2670
	2.0	1.0000	1.0000	1.0000	0.9995	0.9840	0.8575	0.4145
Pooled	0.0	0.0015	0.0070	0.0155	0.0445	0.0990	0.1535	0.2230
	0.5	0.2490	0.2535	0.2705	0.2835	0.2515	0.2680	0.2780
	1.0	0.9465	0.9040	0.8670	0.7555	0.6615	0.5320	0.3655
	1.5	1.0000	0.9995	0.9960	0.9845	0.9185	0.7915	0.4870
	2.0	1.0000	1.0000	1.0000	0.9995	0.9925	0.9220	0.6340
Satterthwaite	0.0	0.0445	0.0520	0.0450	0.0465	0.0465	0.0490	0.0415
	0.5	0.7000	0.4960	0.3815	0.2645	0.1540	0.1130	0.0660
	1.0	0.9985	0.9730	0.9165	0.7225	0.4935	0.2875	0.1170
	1.5	1.0000	1.0000	0.9990	0.9745	0.8250	0.5370	0.1685
	2.0	1.0000	1.0000	1.0000	0.9995	0.9650	0.7740	0.2860

Table A27

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 30$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0050	0.0045	0.0045	0.0065	0.0090	0.0075	0.0075
	0.5	0.3875	0.2020	0.1280	0.0765	0.0330	0.0260	0.0125
	1.0	0.9810	0.8600	0.6650	0.3830	0.2015	0.0965	0.0285
	1.5	1.0000	0.9985	0.9755	0.8255	0.5090	0.2380	0.0465
	2.0	1.0000	1.0000	1.0000	0.9730	0.8110	0.4645	0.0895
Efron & Tibshirani	0.0	0.0065	0.0080	0.0055	0.0075	0.0115	0.0070	0.0050
	0.5	0.4140	0.2420	0.1605	0.0875	0.0385	0.0235	0.0095
	1.0	0.9850	0.8885	0.7100	0.4010	0.2080	0.0800	0.0200
	1.5	1.0000	1.0000	0.9805	0.8255	0.4880	0.2100	0.0285
	2.0	1.0000	1.0000	1.0000	0.9695	0.7795	0.3960	0.0685
Good	0.0	0.0150	0.0140	0.0160	0.0200	0.0285	0.0365	0.0350
	0.5	0.5030	0.3345	0.2595	0.1780	0.1000	0.0810	0.0505
	1.0	0.9935	0.9375	0.8445	0.6145	0.4070	0.2285	0.0975
	1.5	1.0000	1.0000	0.9945	0.9465	0.7590	0.4685	0.1465
	2.0	1.0000	1.0000	1.0000	0.9970	0.9405	0.7120	0.2400
Pooled	0.0	<.0005	0.0010	0.0015	0.0100	0.0275	0.0620	0.1055
	0.5	0.0500	0.0590	0.0885	0.1135	0.1050	0.1415	0.1540
	1.0	0.7180	0.6575	0.6020	0.5195	0.4360	0.3525	0.2330
	1.5	0.9975	0.9895	0.9690	0.9235	0.8030	0.6200	0.3270
	2.0	1.0000	1.0000	1.0000	0.9965	0.9655	0.8440	0.4835
Satterthwaite	0.0	0.0075	0.0090	0.0065	0.0080	0.0140	0.0100	0.0080
	0.5	0.4270	0.2535	0.1725	0.1045	0.0440	0.0305	0.0130
	1.0	0.9865	0.9010	0.7405	0.4405	0.2390	0.1060	0.0310
	1.5	1.0000	1.0000	0.9860	0.8645	0.5500	0.2580	0.0510
	2.0	1.0000	1.0000	1.0000	0.9835	0.8390	0.4885	0.0965

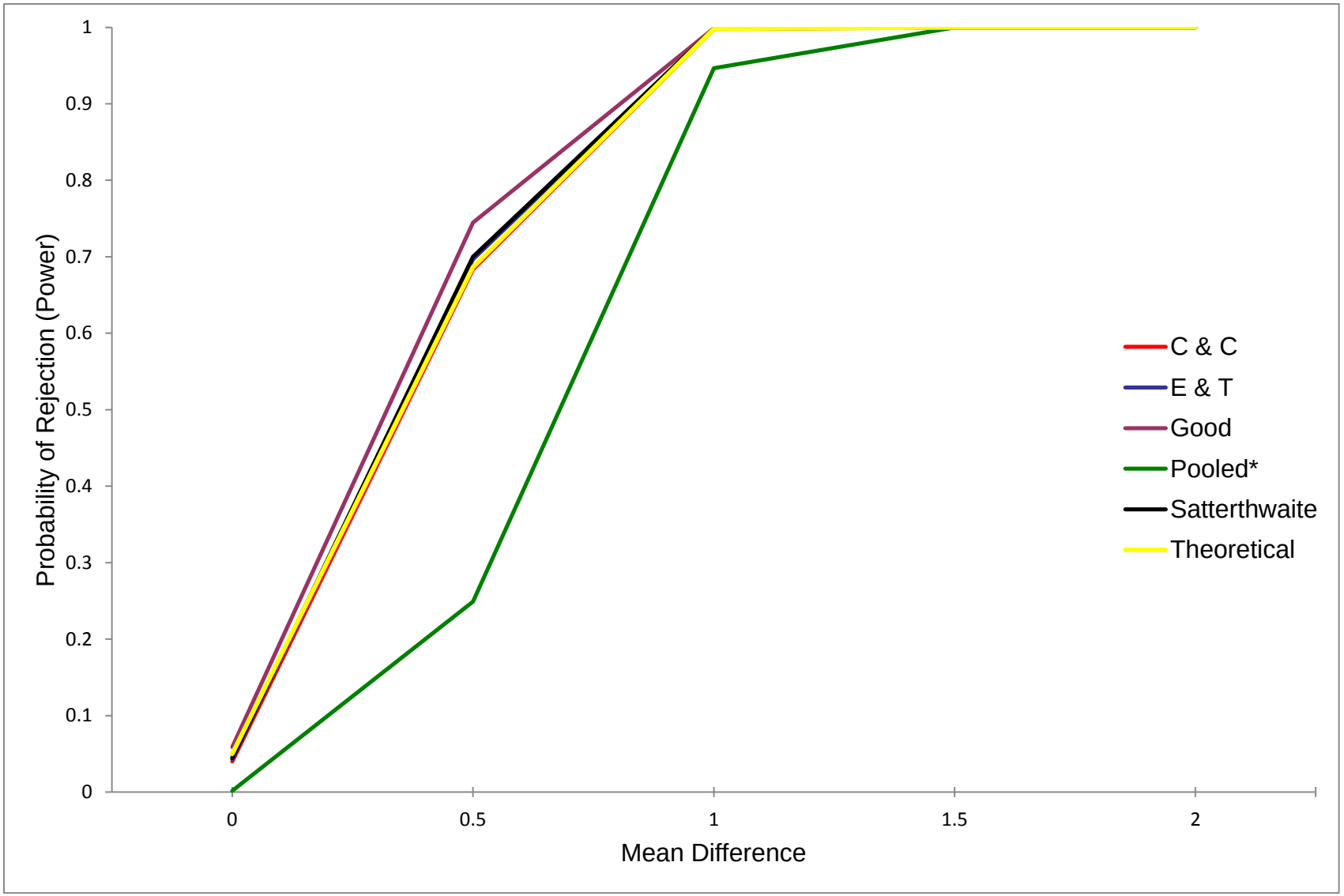

Figure A29. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

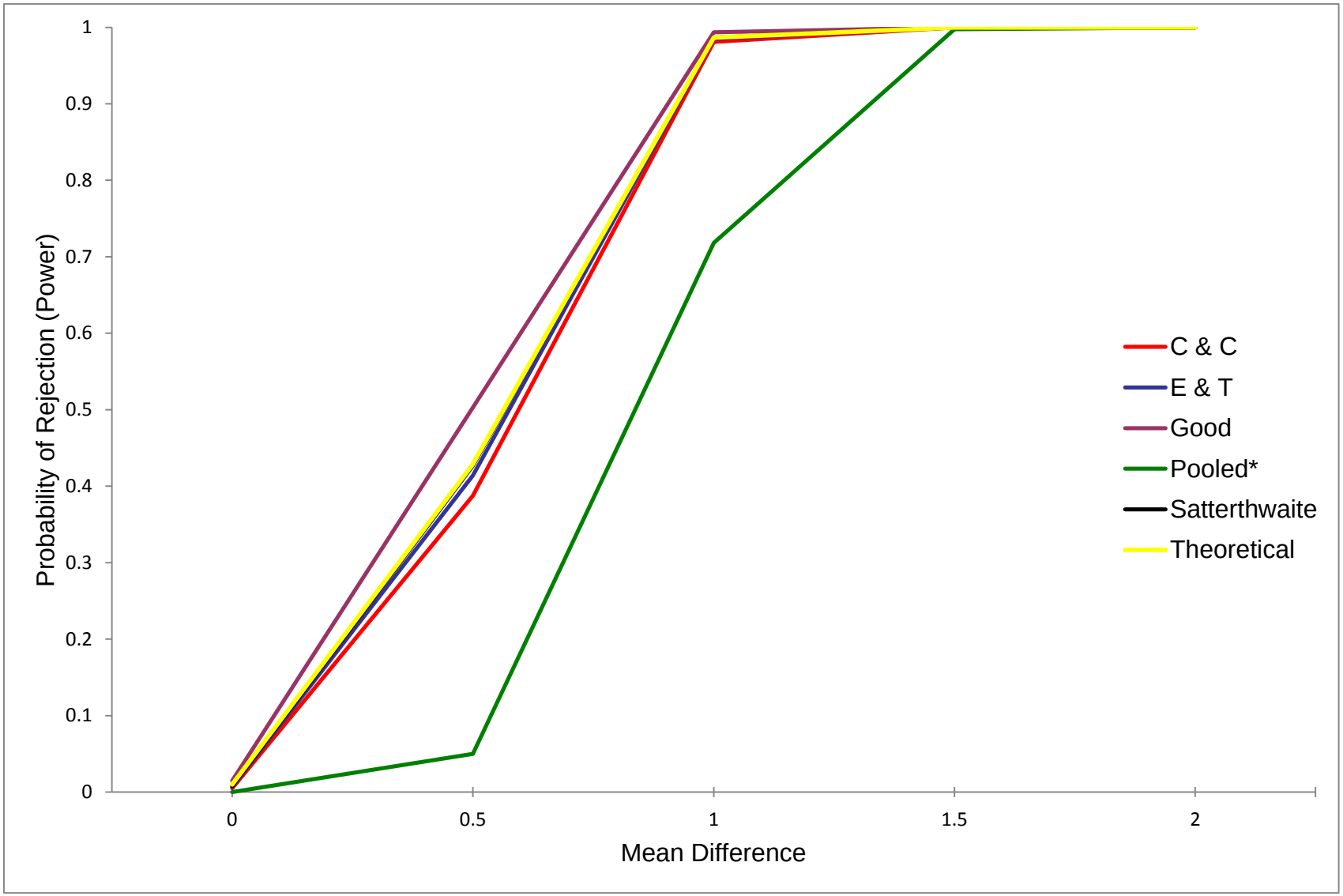

Figure A30. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

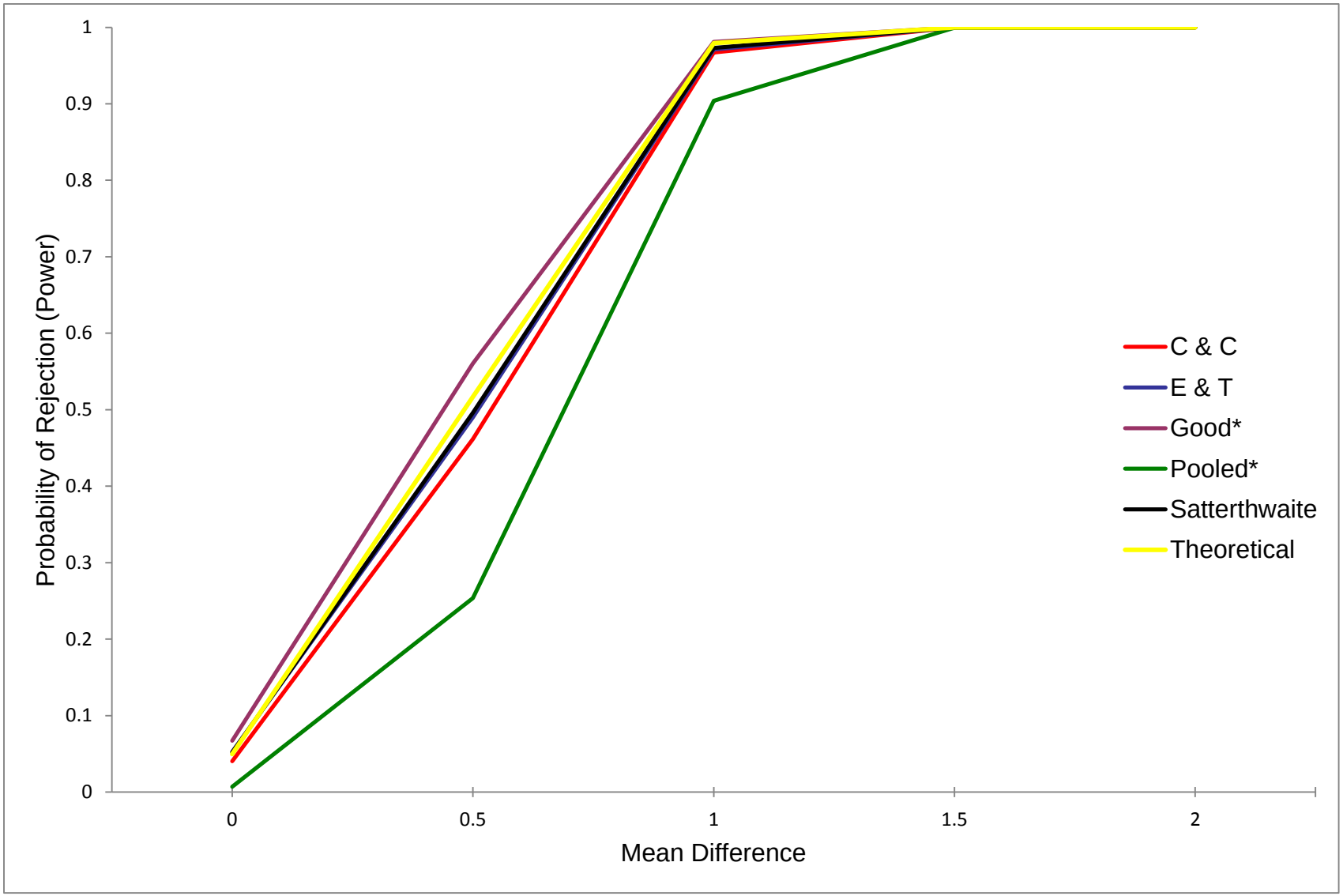

Figure A31. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

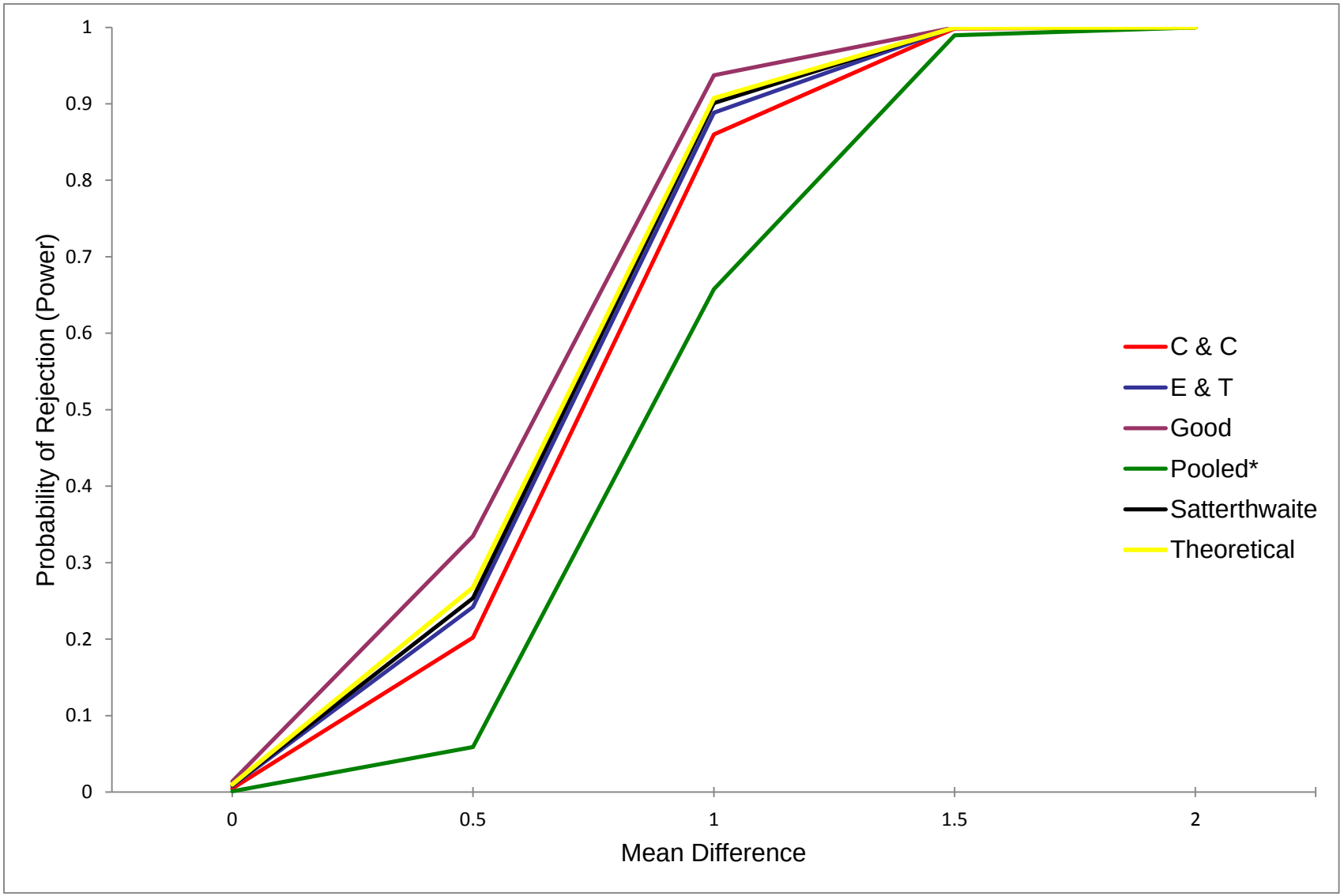

Figure A32. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

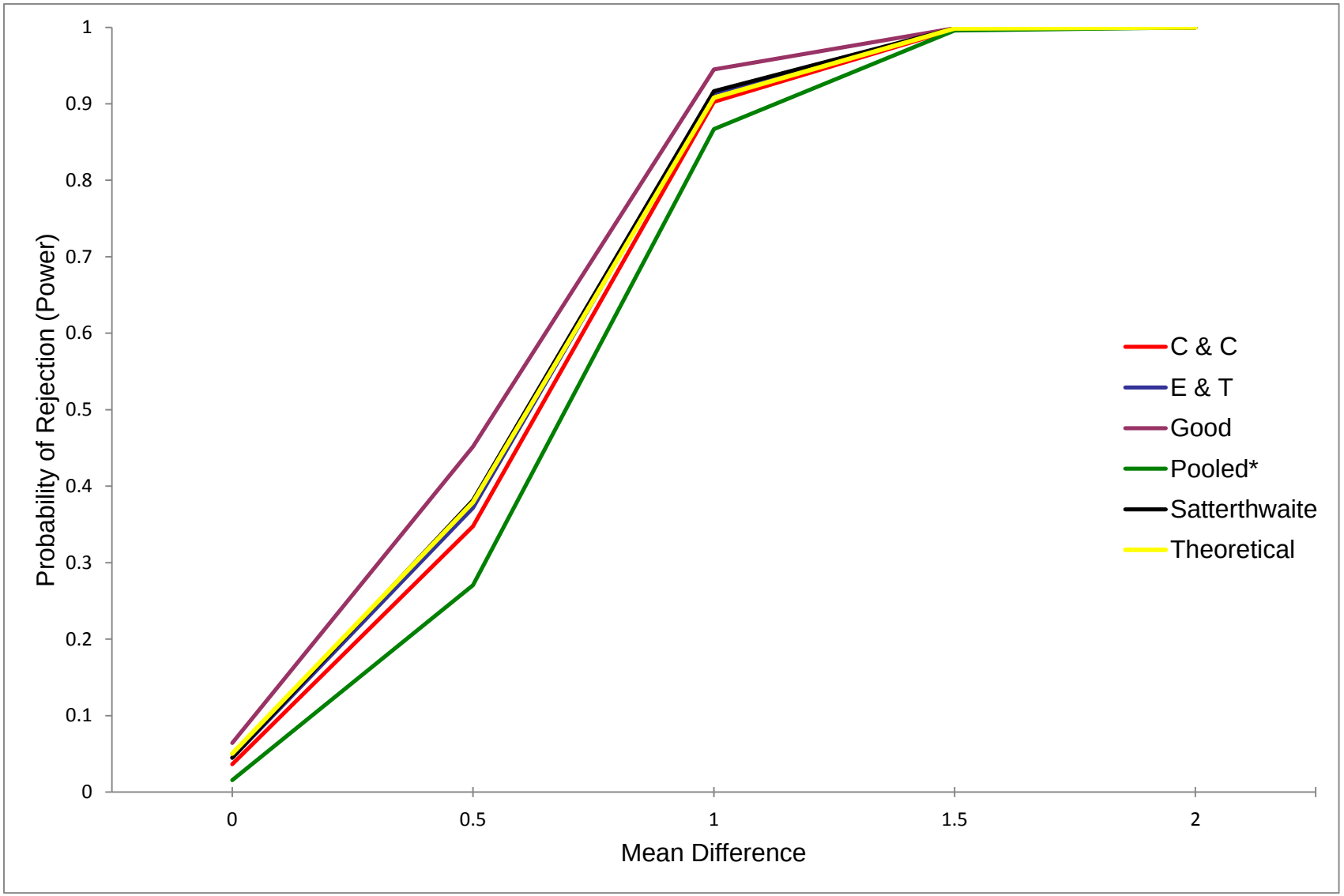

Figure A33. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

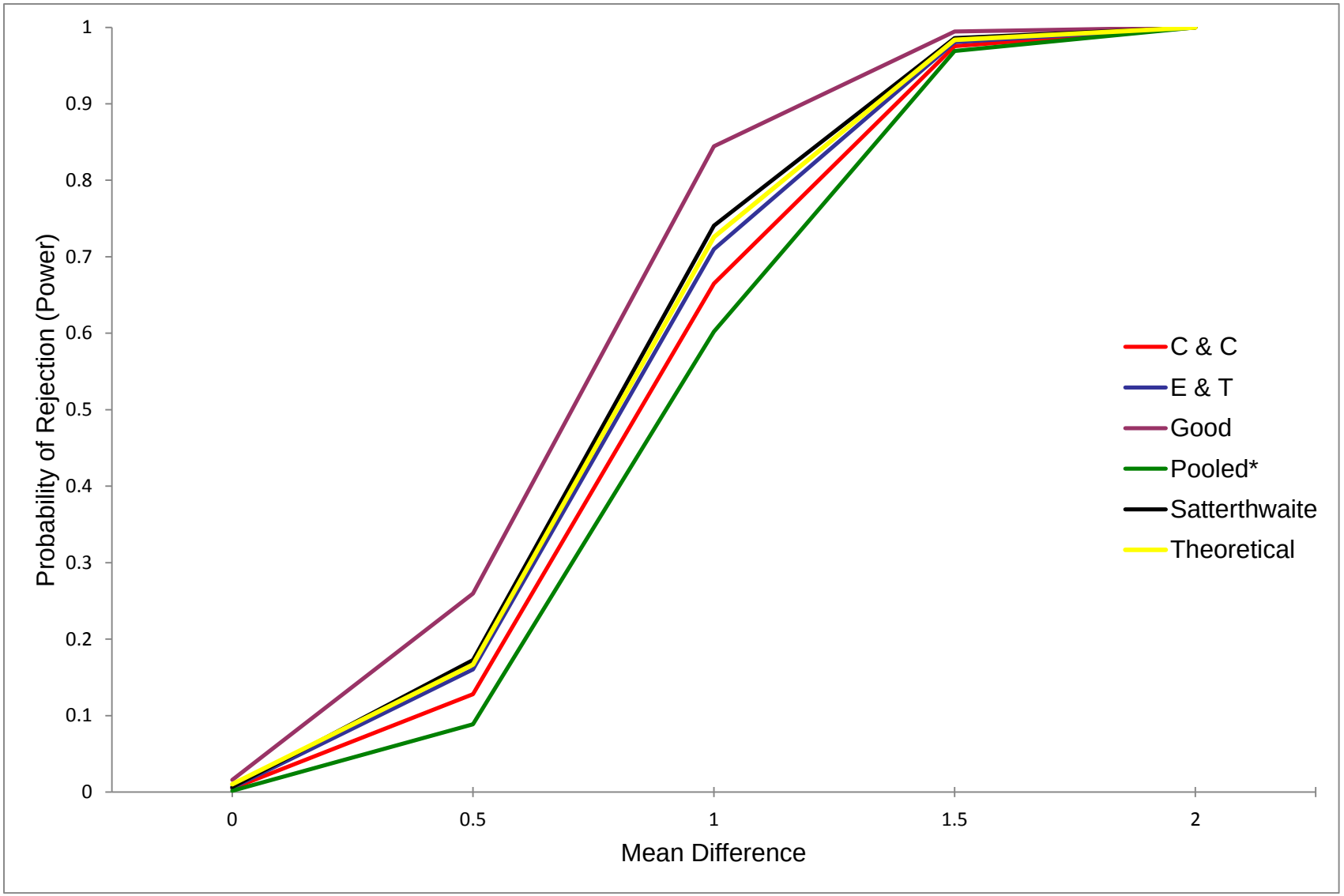

Figure A34. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

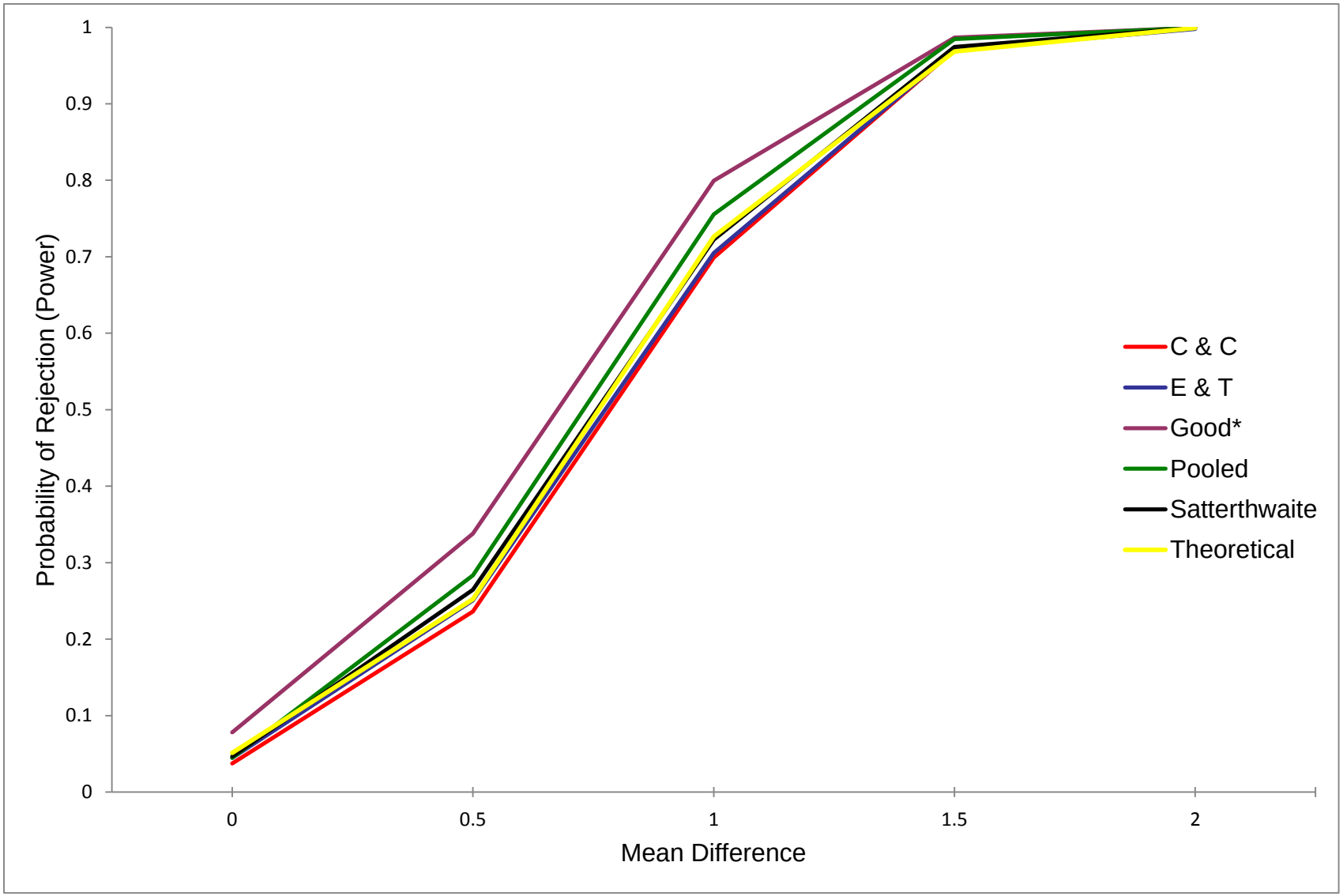

Figure A35. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

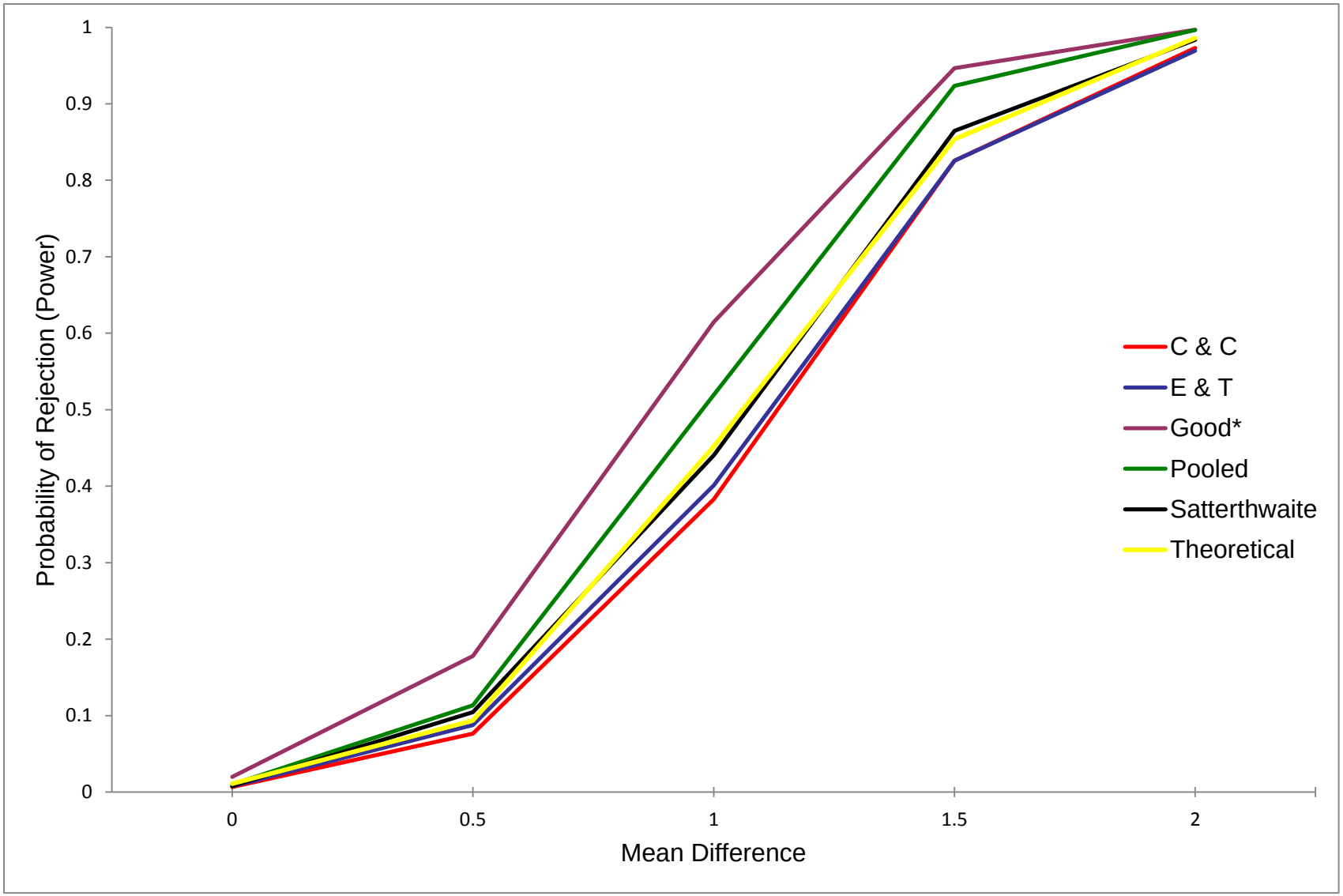

Figure A36. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

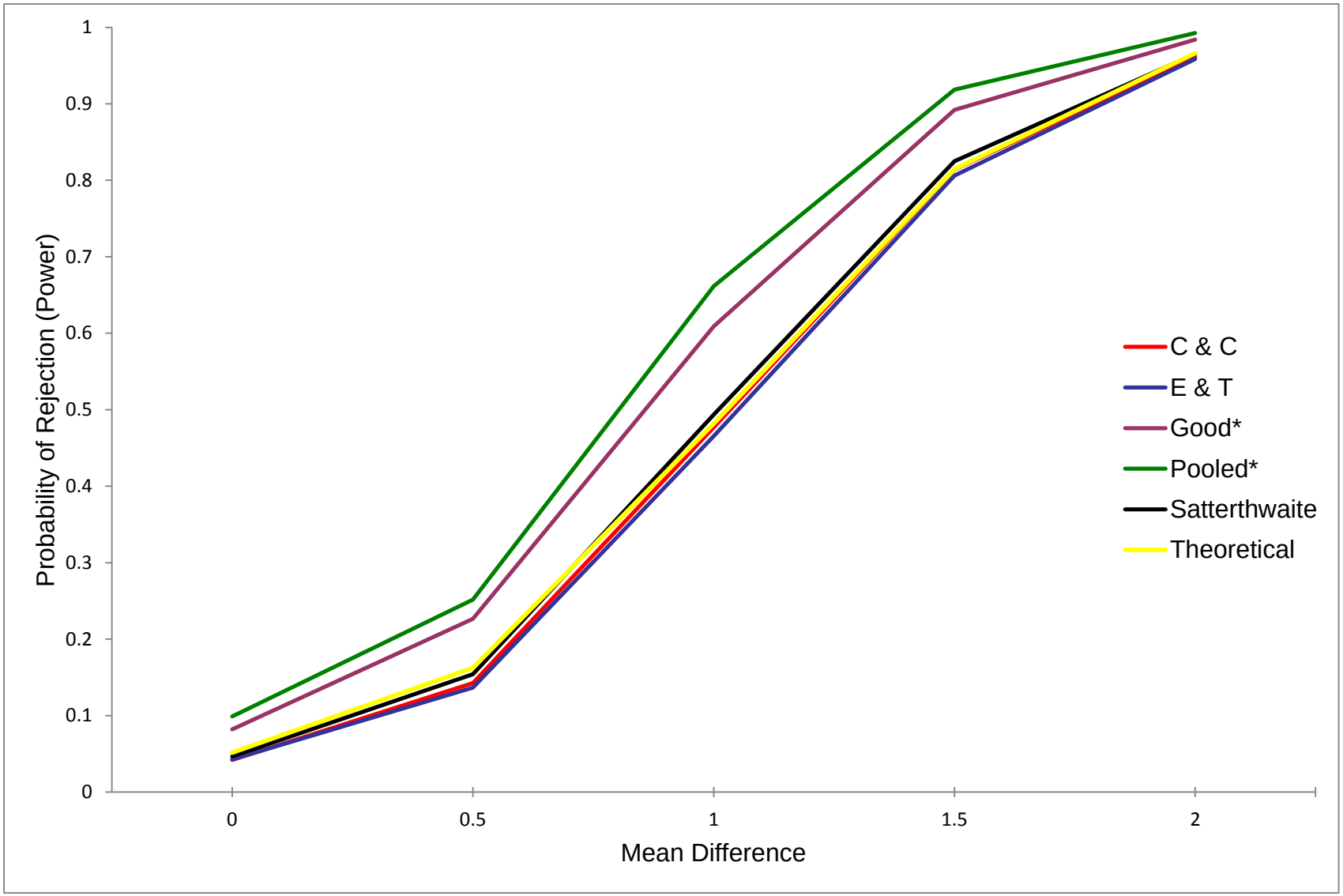

Figure A37. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

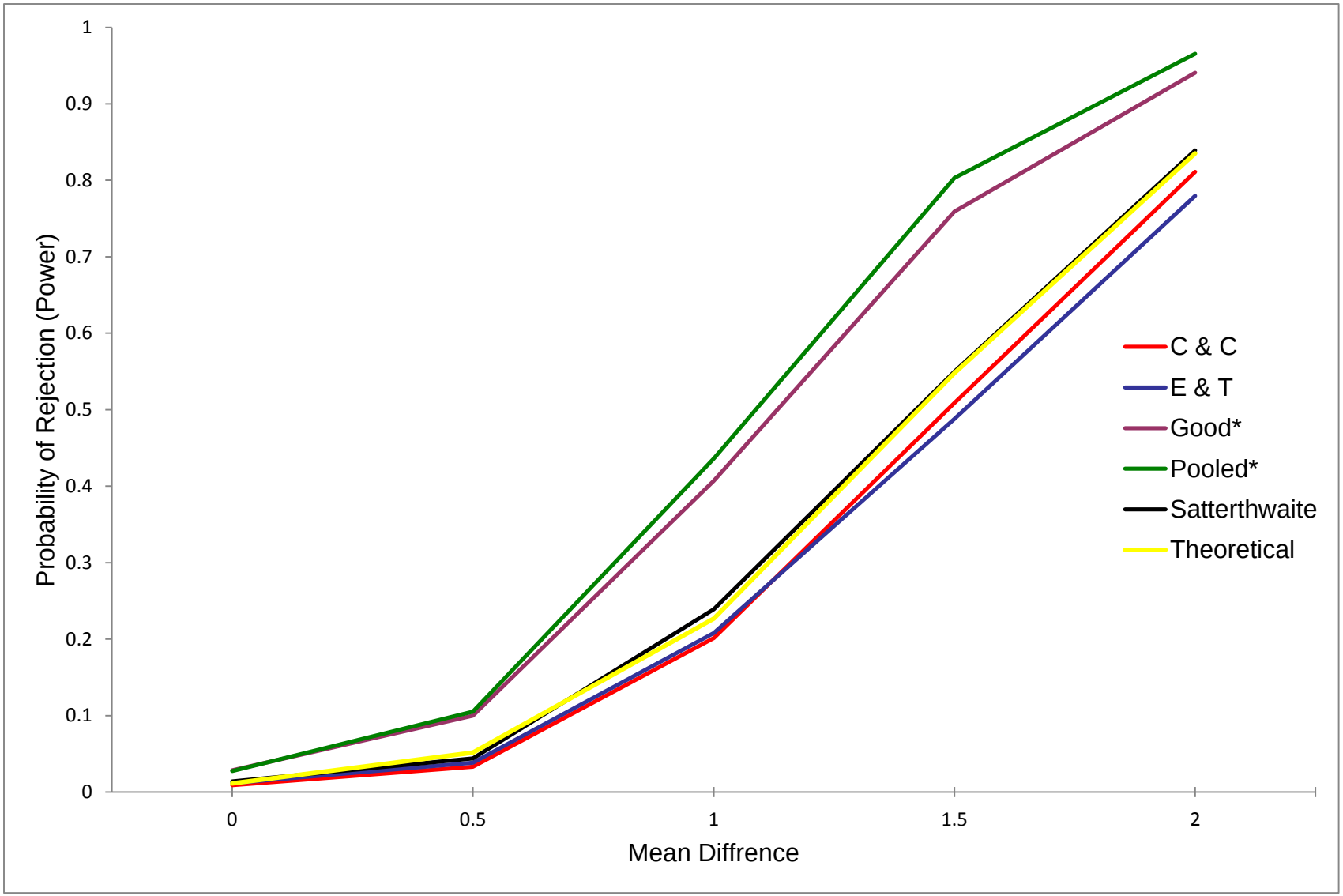

Figure A38. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

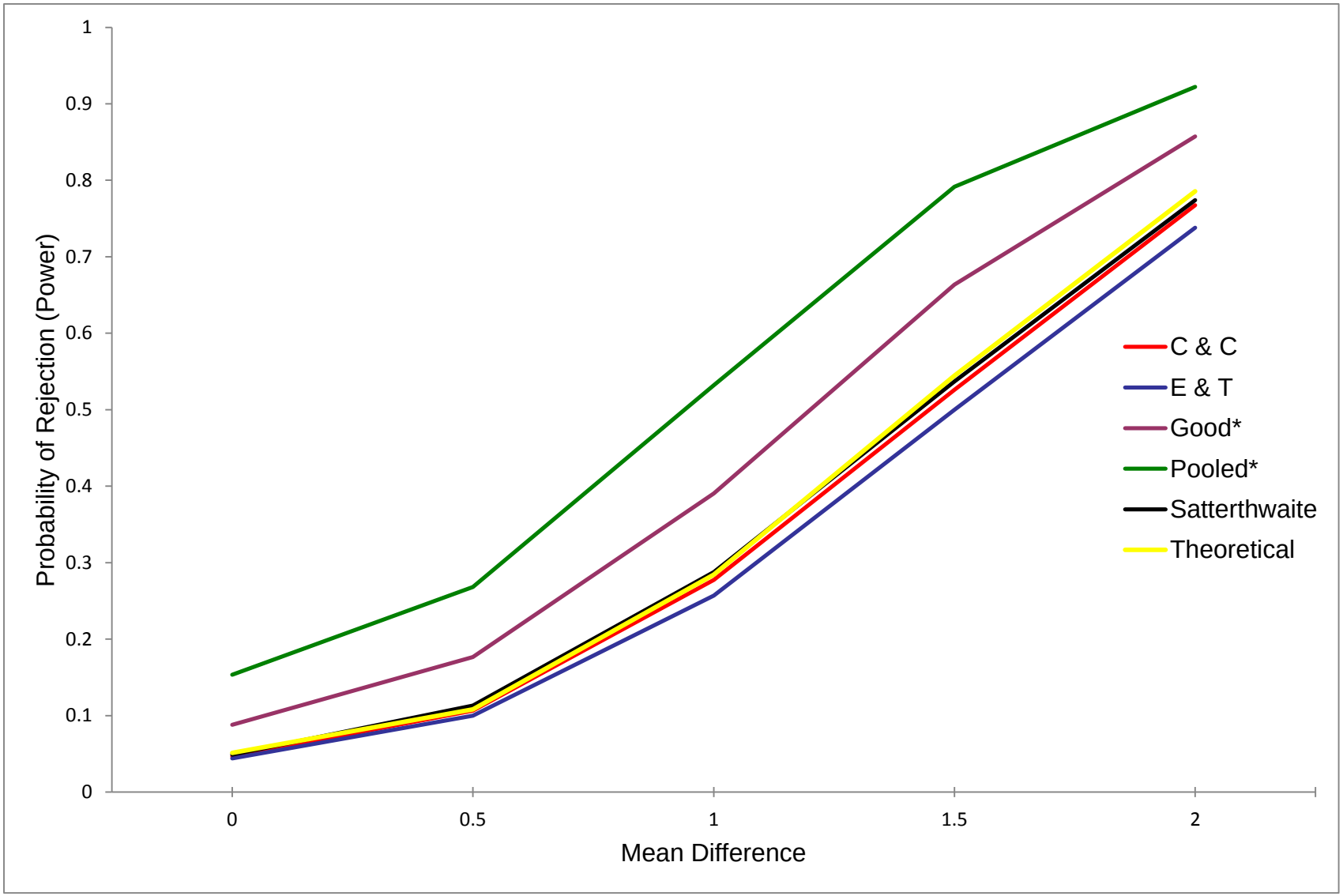

Figure A39. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

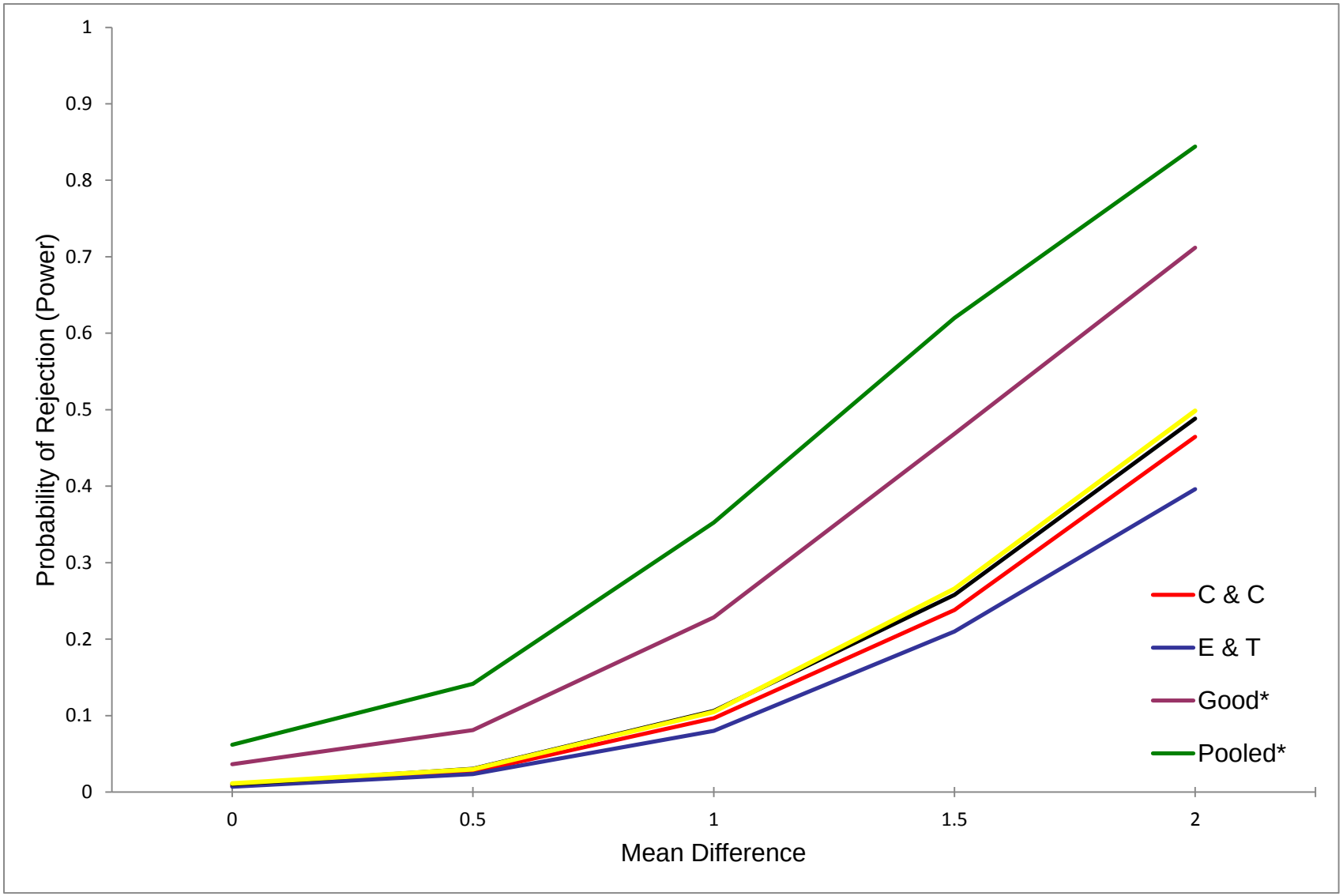

Figure A40. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

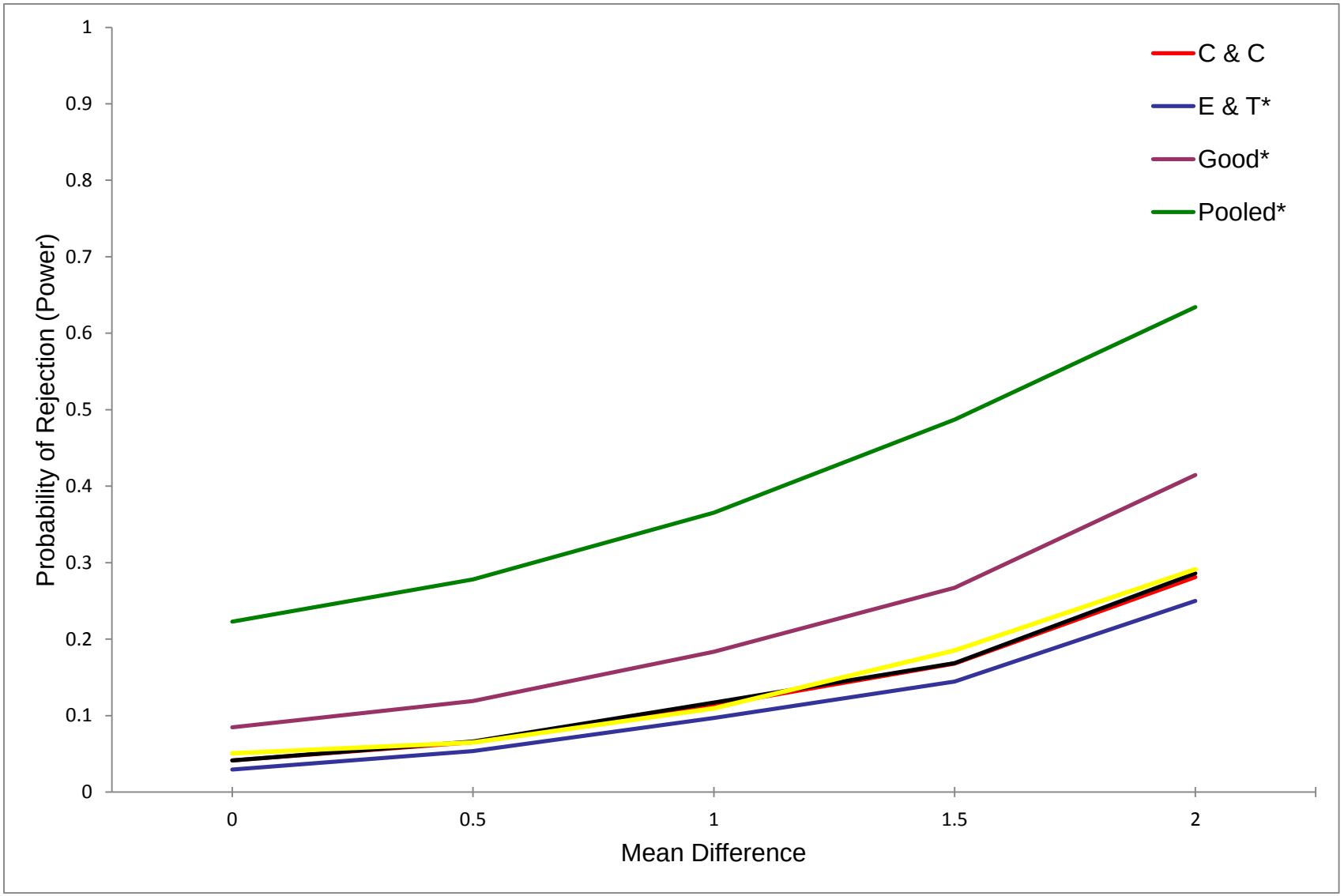

Figure A41. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

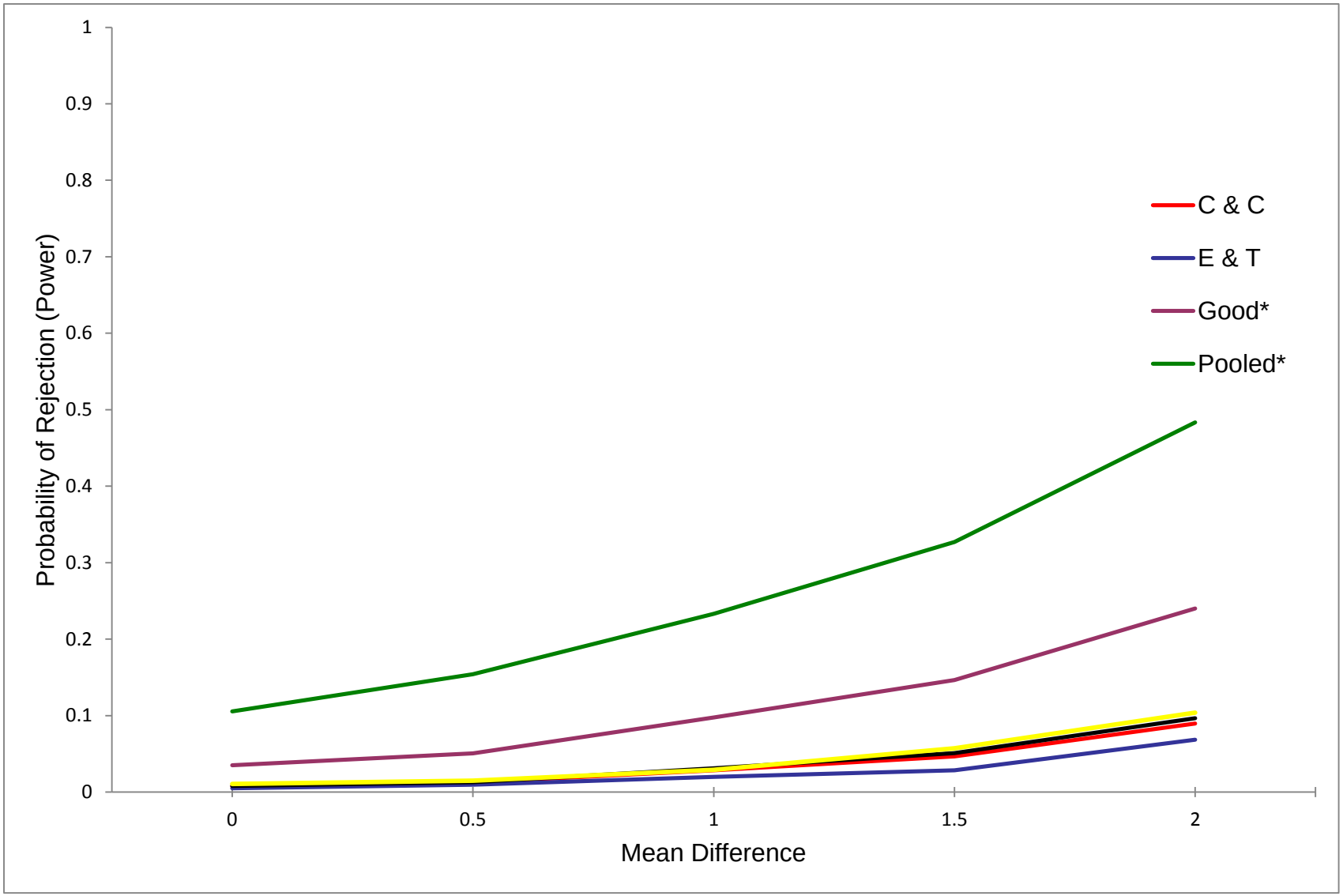

Figure A42. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 30$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 5.0 (i.e., $n_1 = 10$, $n_2 = 50$)

Table A28

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0360	0.0060
E & T	0.0430	0.0085
Good	0.0560	0.0145
Pooled	0.0005*	<.0005*
Satterthwaite	0.0435	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A29

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0345	0.0075
E & T	0.0435	0.0130
Good	0.0610	0.0210*
Pooled	0.0040*	0.0005*
Satterthwaite	0.0450	0.0125

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A30

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0435	0.0105
E & T	0.0505	0.0140
Good	0.0760*	0.0265*
Pooled	0.0155*	0.0010*
Satterthwaite	0.0530	0.0155

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A31

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0480	0.0100
E & T	0.0505	0.0105
Good	0.0885*	0.0290*
Pooled	0.0565	0.0125
Satterthwaite	0.0530	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A32

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0410	0.0085
E & T	0.0385	0.0080
Good	0.0840*	0.0280*
Pooled	0.1090*	0.0380*
Satterthwaite	0.0475	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A33

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0475	0.0085
E & T	0.0435	0.0080
Good	0.0980*	0.0385*
Pooled	0.2050*	0.1000*
Satterthwaite	0.0490	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A34

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0585	0.0140
E & T	0.0515	0.0090
Good	0.1190*	0.0475*
Pooled	0.3320*	0.2140*
Satterthwaite	0.0590	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table A35

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0360	0.0345	0.0435	0.0480	0.0410	0.0475	0.0585
	0.5	0.8490	0.5945	0.4175	0.2510	0.1535	0.0985	0.0645
	1.0	1.0000	0.9935	0.9320	0.7520	0.5025	0.3030	0.1210
	1.5	1.0000	1.0000	0.9990	0.9810	0.8115	0.5415	0.1885
	2.0	1.0000	1.0000	1.0000	1.0000	0.9705	0.7920	0.2870
Efron & Tibshirani	0.0	0.0430	0.0435	0.0505	0.0505	0.0385	0.0435	0.0515
	0.5	0.8610	0.6250	0.4305	0.2585	0.1445	0.0880	0.0515
	1.0	1.0000	0.9950	0.9325	0.7445	0.4815	0.2830	0.1000
	1.5	1.0000	1.0000	0.9990	0.9760	0.7885	0.5065	0.1660
	2.0	1.0000	1.0000	1.0000	1.0000	0.9615	0.7545	0.2400
Good	0.0	0.0560	0.0610	0.0760	0.0885	0.0840	0.0980	0.1190
	0.5	0.8760	0.6940	0.5270	0.3525	0.2490	0.1660	0.1100
	1.0	1.0000	0.9980	0.9660	0.8325	0.6245	0.4250	0.1995
	1.5	1.0000	1.0000	0.9995	0.9920	0.8930	0.6720	0.2860
	2.0	1.0000	1.0000	1.0000	1.0000	0.9890	0.8790	0.4175
Pooled	0.0	0.0005	0.0040	0.0155	0.0565	0.1090	0.2050	0.3320
	0.5	0.2310	0.2595	0.2910	0.2860	0.3100	0.3220	0.3675
	1.0	0.9800	0.9470	0.8850	0.8120	0.6985	0.6090	0.4845
	1.5	1.0000	1.0000	0.9985	0.9915	0.9420	0.8455	0.5760
	2.0	1.0000	1.0000	1.0000	1.0000	0.9970	0.9585	0.7280
Satterthwaite	0.0	0.0435	0.0450	0.0530	0.0530	0.0475	0.0490	0.0590
	0.5	0.8620	0.6345	0.4450	0.2705	0.1615	0.1040	0.0655
	1.0	1.0000	0.9950	0.9380	0.7650	0.5100	0.3100	0.1220
	1.5	1.0000	1.0000	0.9990	0.9835	0.8205	0.5490	0.1905
	2.0	1.0000	1.0000	1.0000	1.0000	0.9740	0.7970	0.2895

Table A36

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 10$, $n_2 = 50$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0060	0.0075	0.0105	0.0100	0.0085	0.0085	0.0140
	0.5	0.6200	0.3075	0.1665	0.0910	0.0465	0.0235	0.0105
	1.0	0.9995	0.9410	0.7355	0.4390	0.2175	0.1070	0.0310
	1.5	1.0000	0.9995	0.9860	0.8570	0.5060	0.2590	0.0615
	2.0	1.0000	1.0000	1.0000	0.9880	0.8140	0.4885	0.1015
Efron & Tibshirani	0.0	0.0085	0.0130	0.0140	0.0105	0.0080	0.0080	0.0090
	0.5	0.6590	0.3600	0.1885	0.0960	0.0370	0.0195	0.0090
	1.0	1.0000	0.9575	0.7530	0.4350	0.2035	0.0820	0.0180
	1.5	1.0000	1.0000	0.9825	0.8310	0.4595	0.2065	0.0370
	2.0	1.0000	1.0000	1.0000	0.9750	0.7480	0.4085	0.0725
Good	0.0	0.0145	0.0210	0.0265	0.0290	0.0280	0.0385	0.0475
	0.5	0.7180	0.4725	0.3255	0.1900	0.1225	0.0790	0.0510
	1.0	1.0000	0.9825	0.8875	0.6755	0.4330	0.2535	0.1065
	1.5	1.0000	1.0000	0.9985	0.9625	0.7665	0.4855	0.1750
	2.0	1.0000	1.0000	1.0000	1.0000	0.9540	0.7420	0.2450
Pooled	0.0	<.0005	0.0005	0.0010	0.0125	0.0380	0.1000	0.2140
	0.5	0.0315	0.0575	0.0860	0.1260	0.1555	0.1755	0.2385
	1.0	0.7805	0.7375	0.6610	0.5945	0.5160	0.4545	0.3355
	1.5	1.0000	0.9965	0.9890	0.9550	0.8560	0.7115	0.4560
	2.0	1.0000	1.0000	1.0000	0.9975	0.9845	0.9090	0.6175
Satterthwaite	0.0	0.0090	0.0125	0.0155	0.0120	0.0095	0.0105	0.0140
	0.5	0.6700	0.3685	0.2005	0.1050	0.0535	0.0270	0.0115
	1.0	1.0000	0.9620	0.7825	0.4905	0.2395	0.1160	0.0315
	1.5	1.0000	1.0000	0.9895	0.8785	0.5355	0.2695	0.0620
	2.0	1.0000	1.0000	1.0000	0.9920	0.8315	0.5055	0.1045

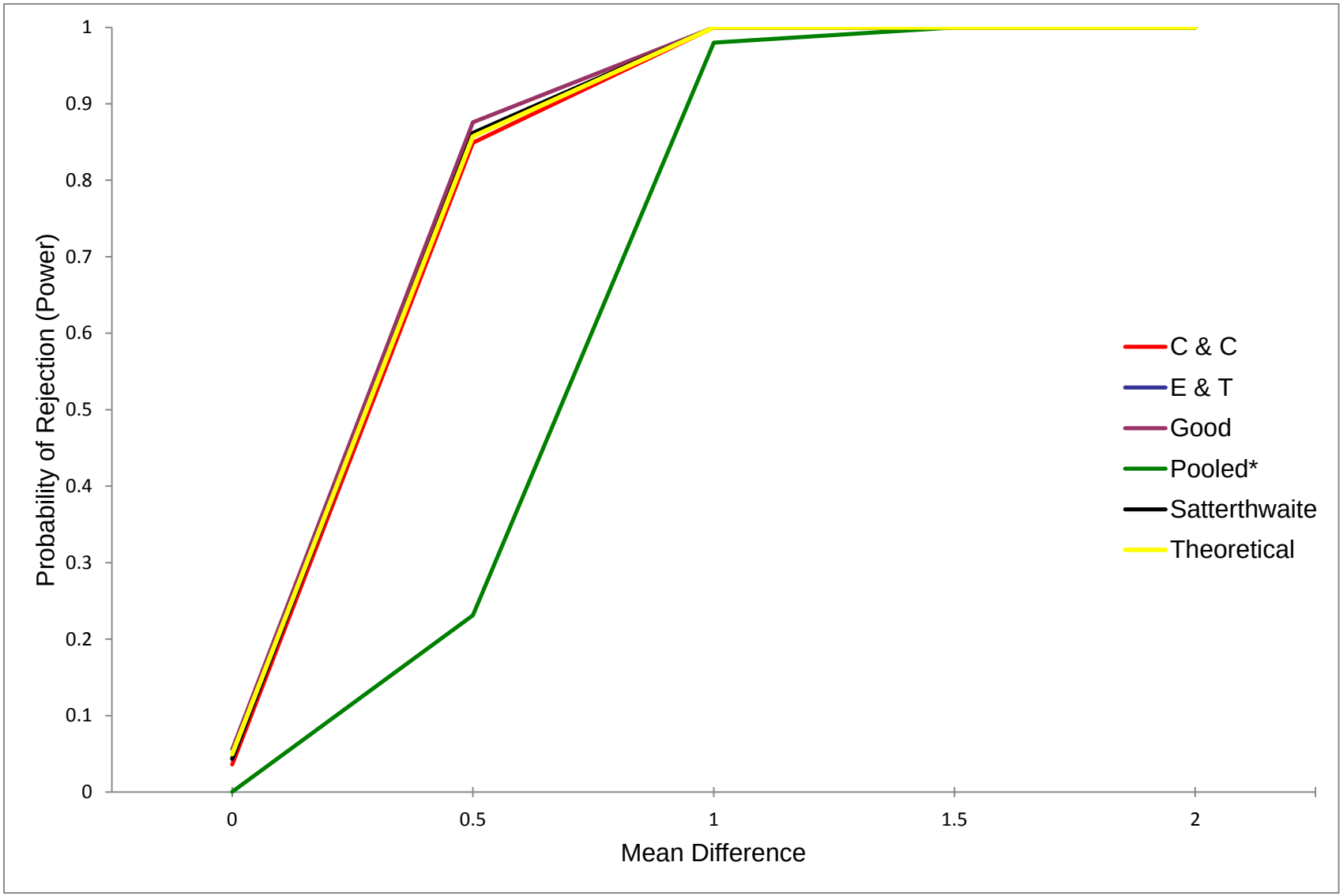

Figure A43. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

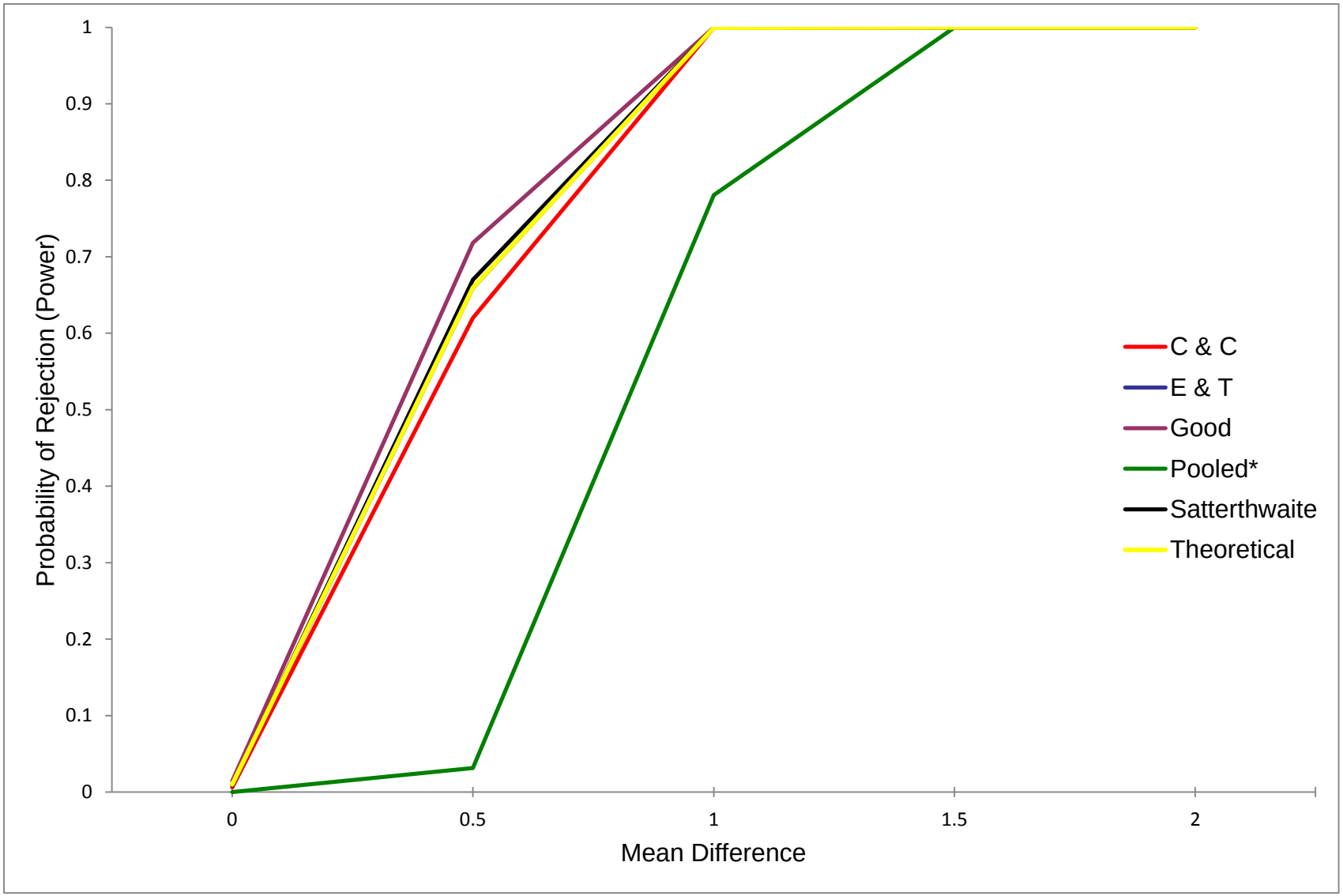

Figure A44. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

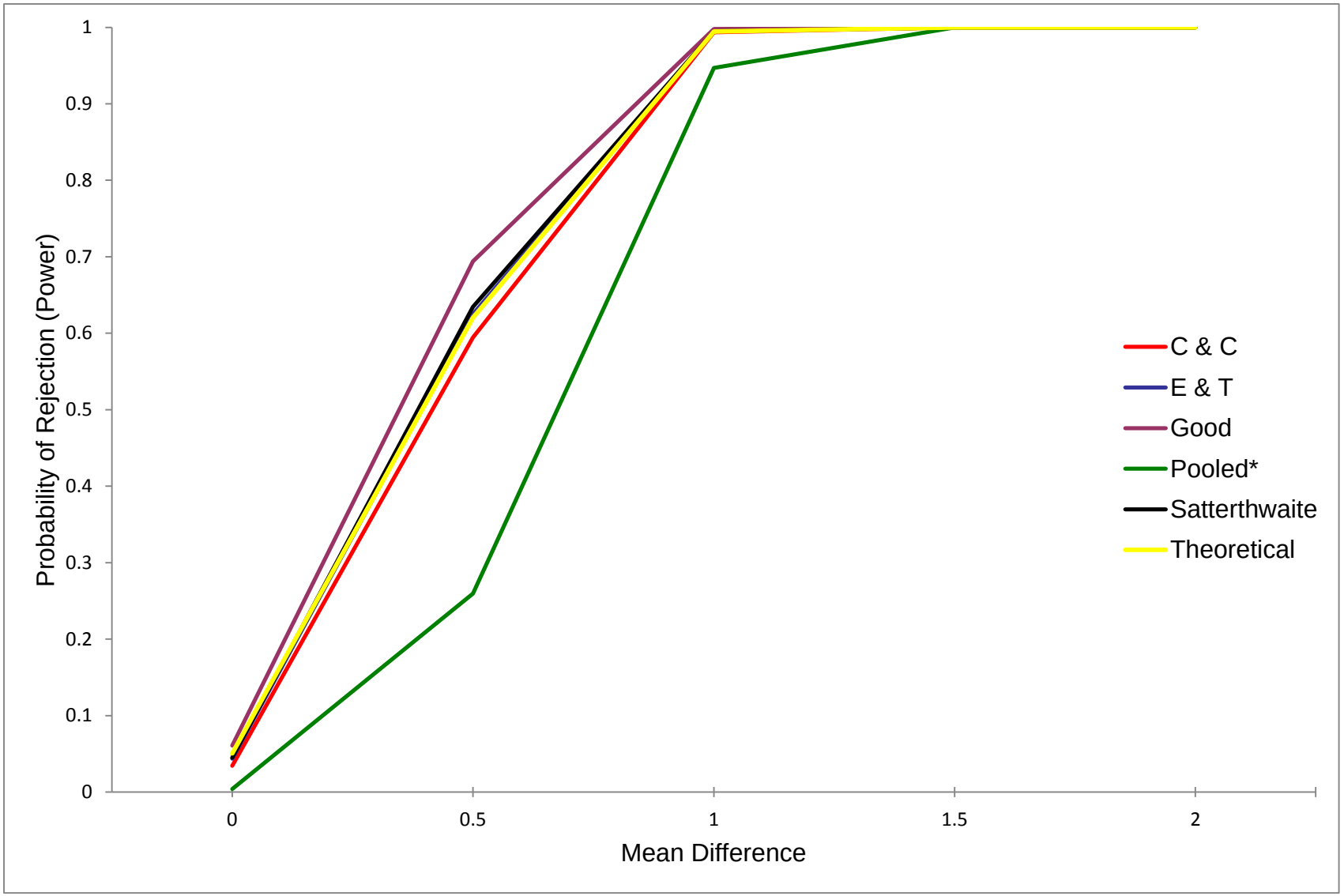

Figure A45. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

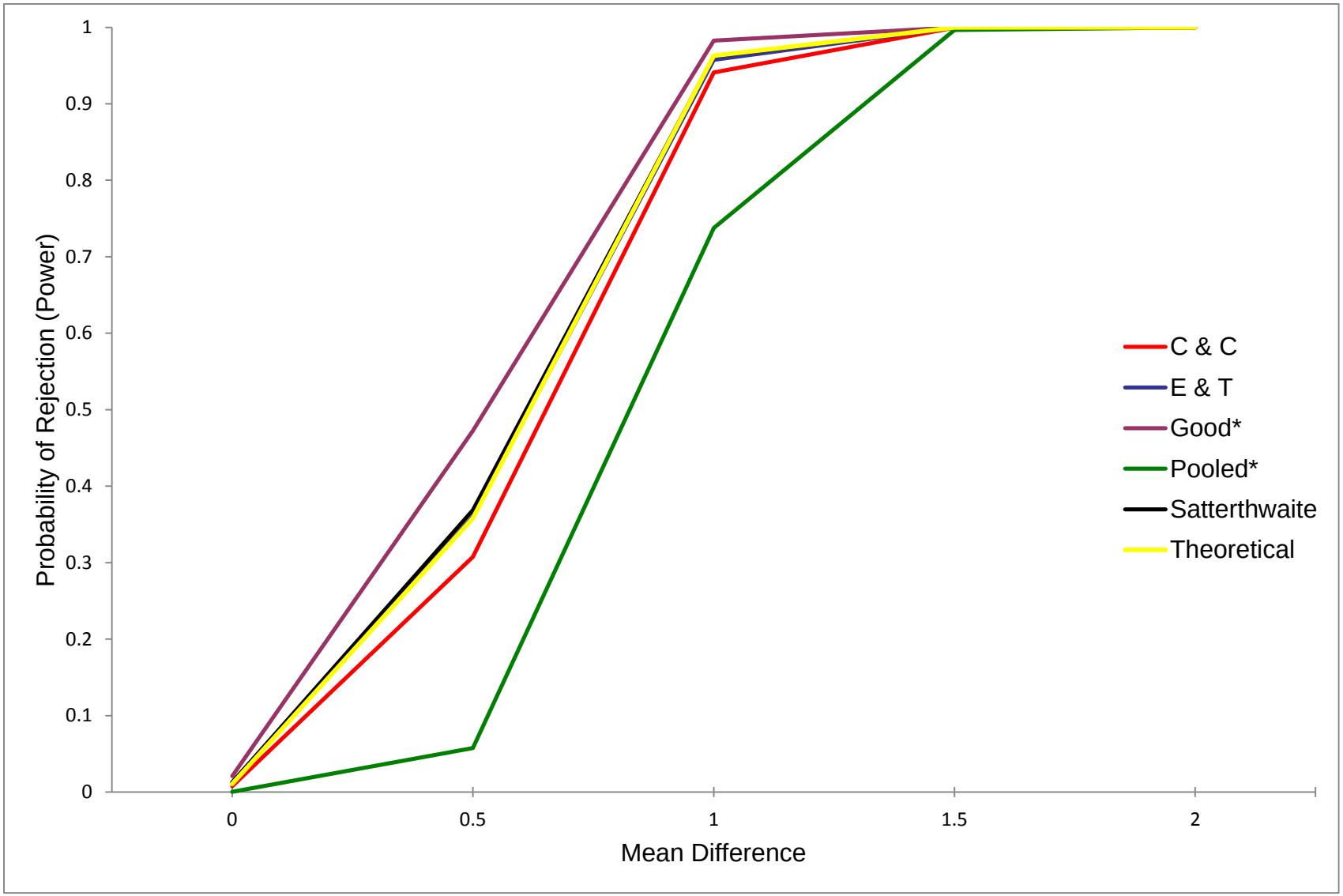

Figure A46. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

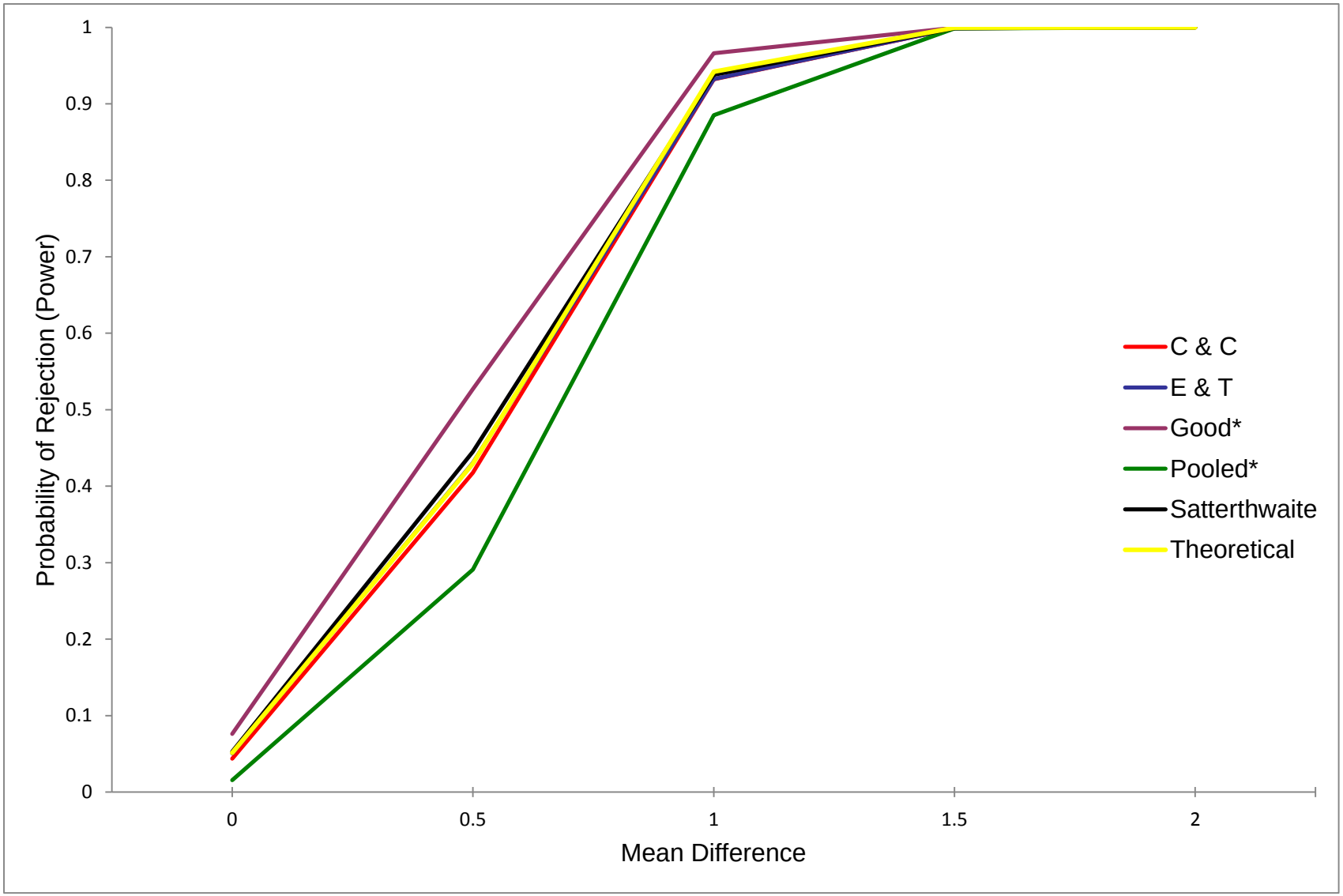

Figure A47. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

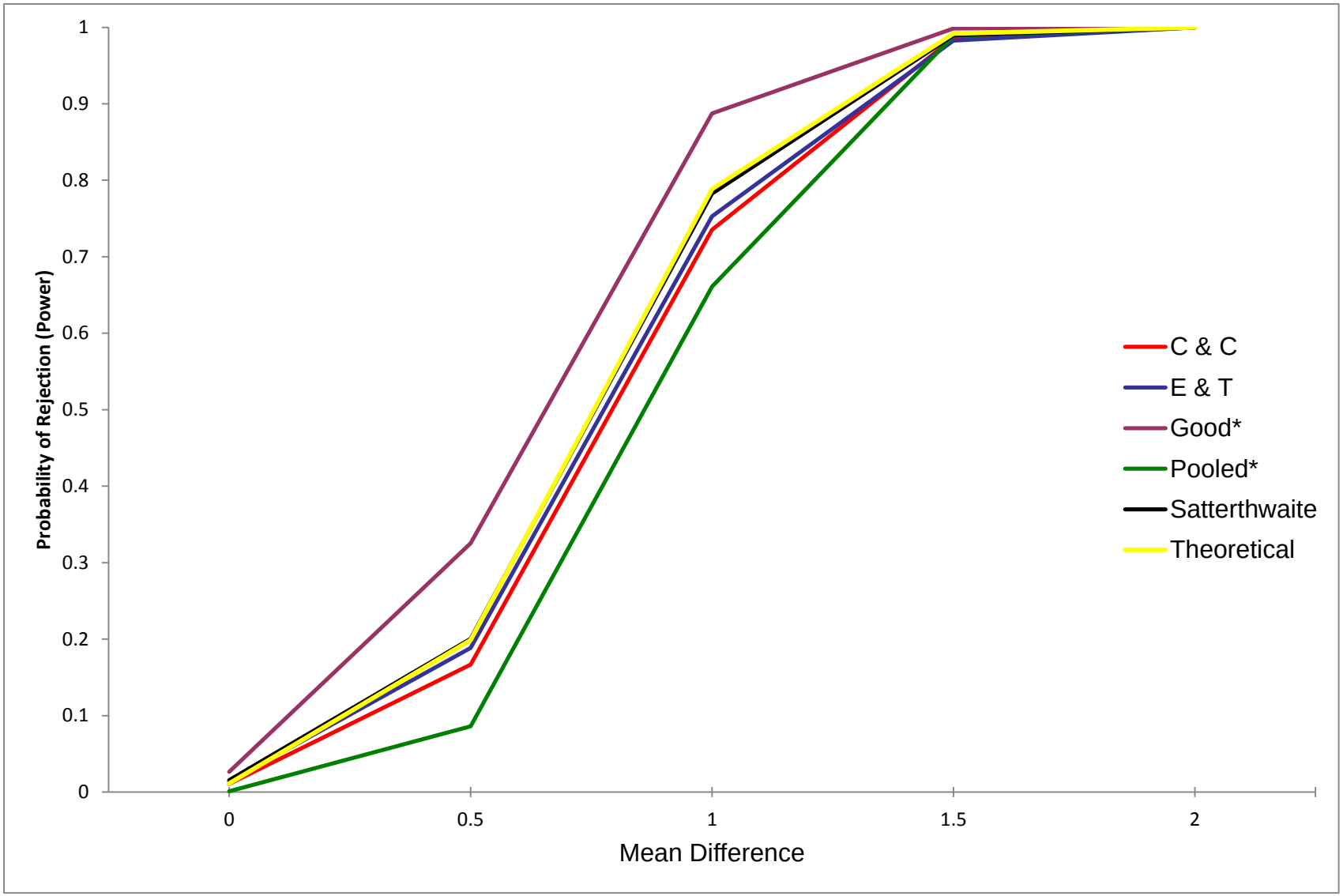

Figure A48. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

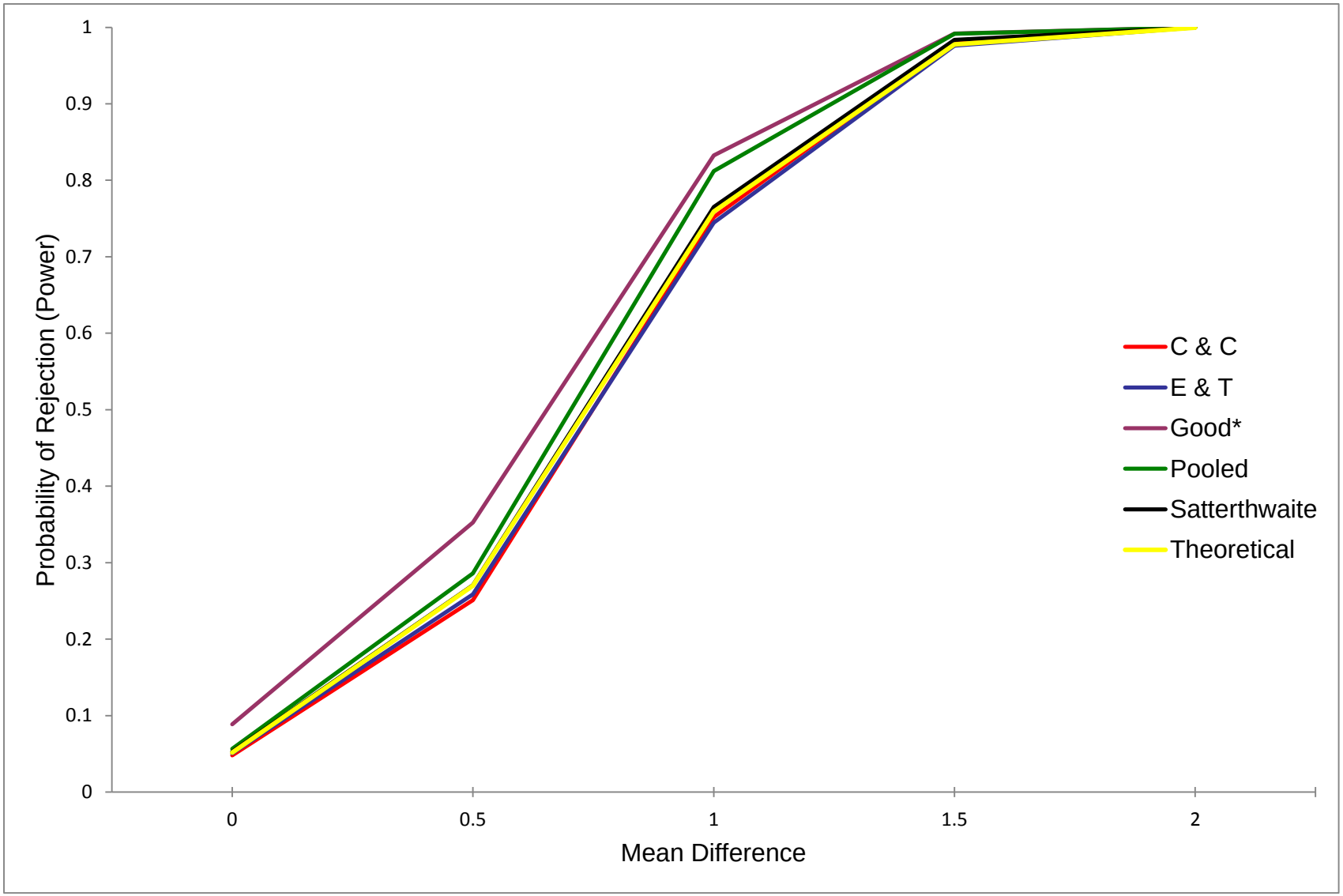

Figure A49. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

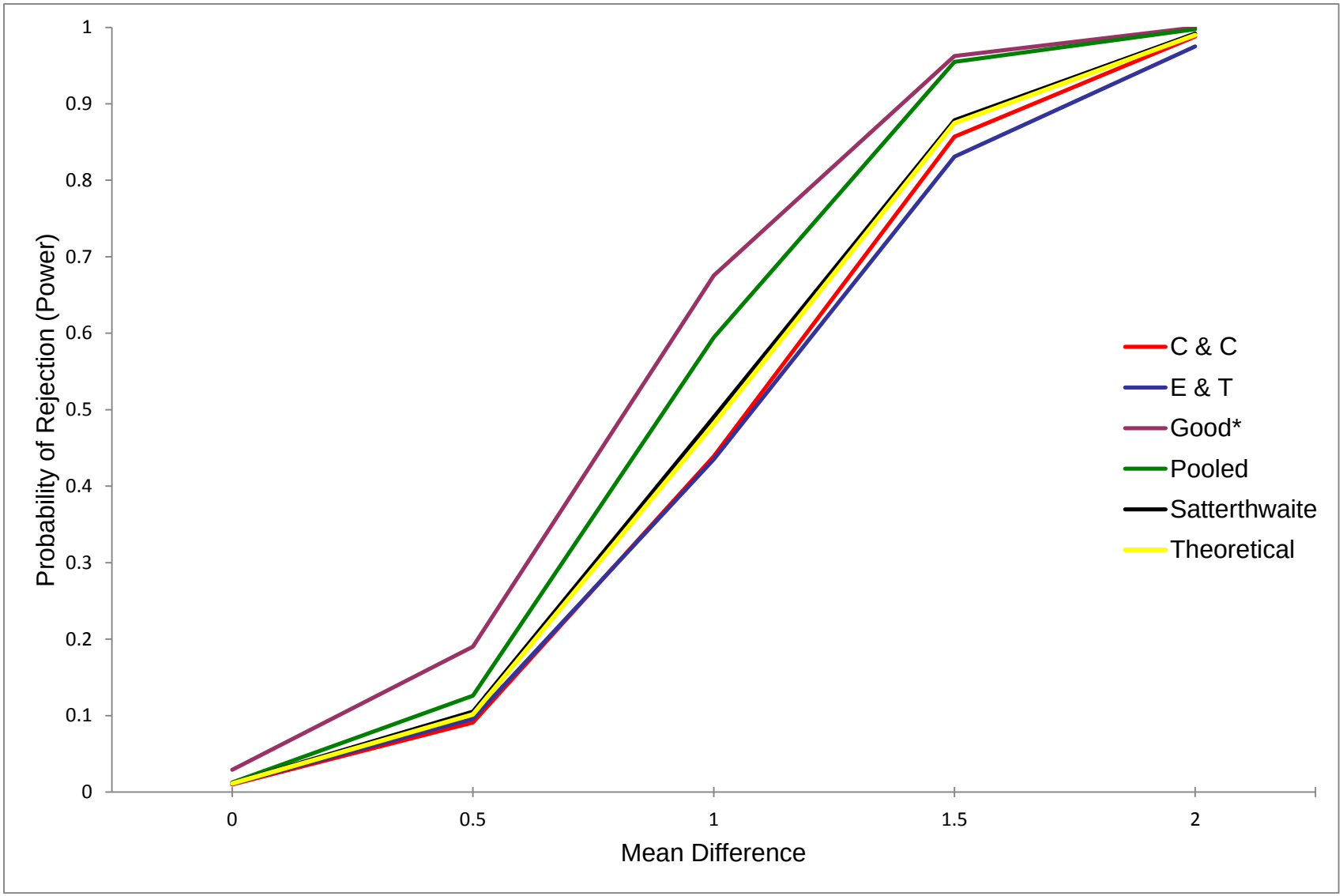

Figure A50. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

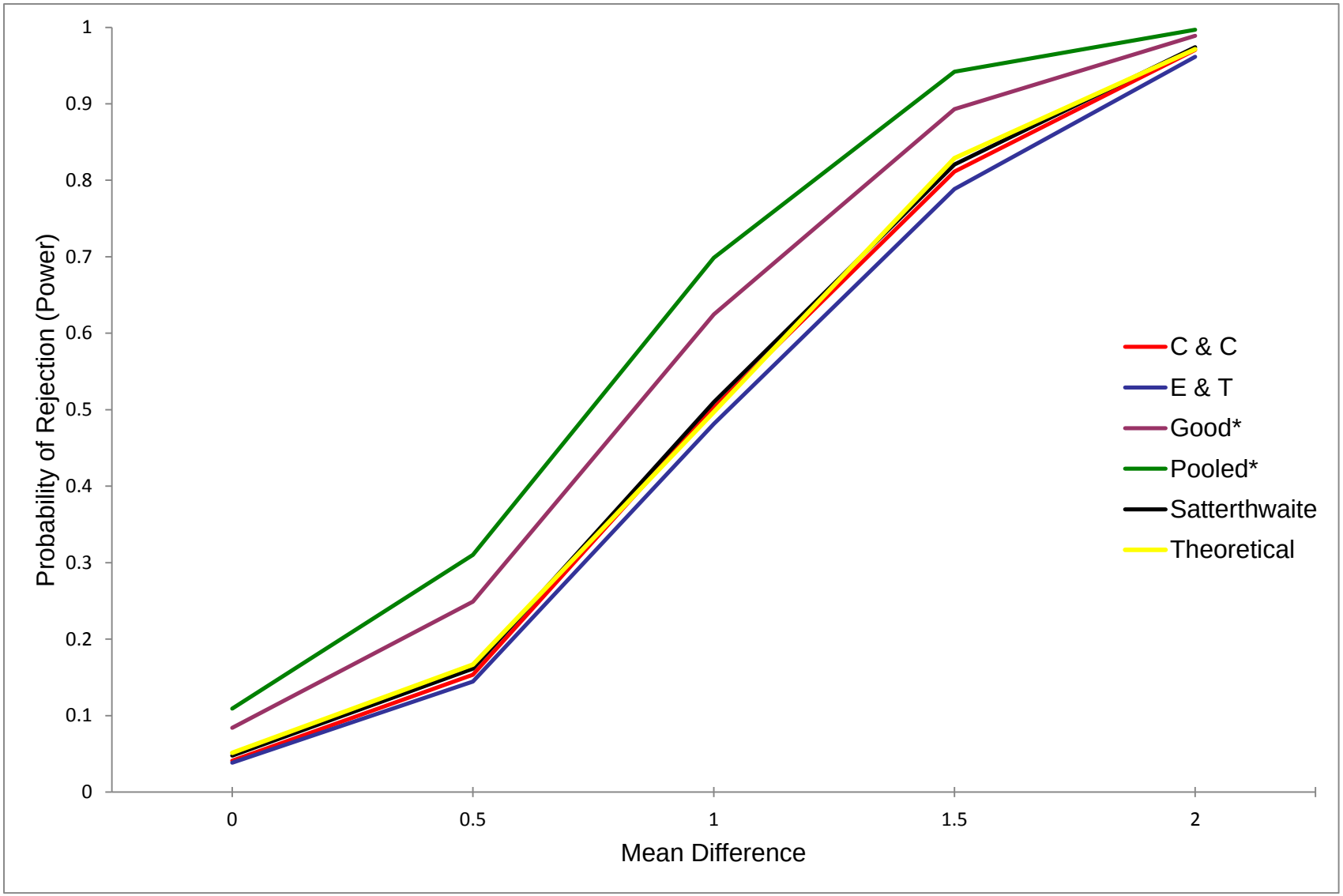

Figure A51. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

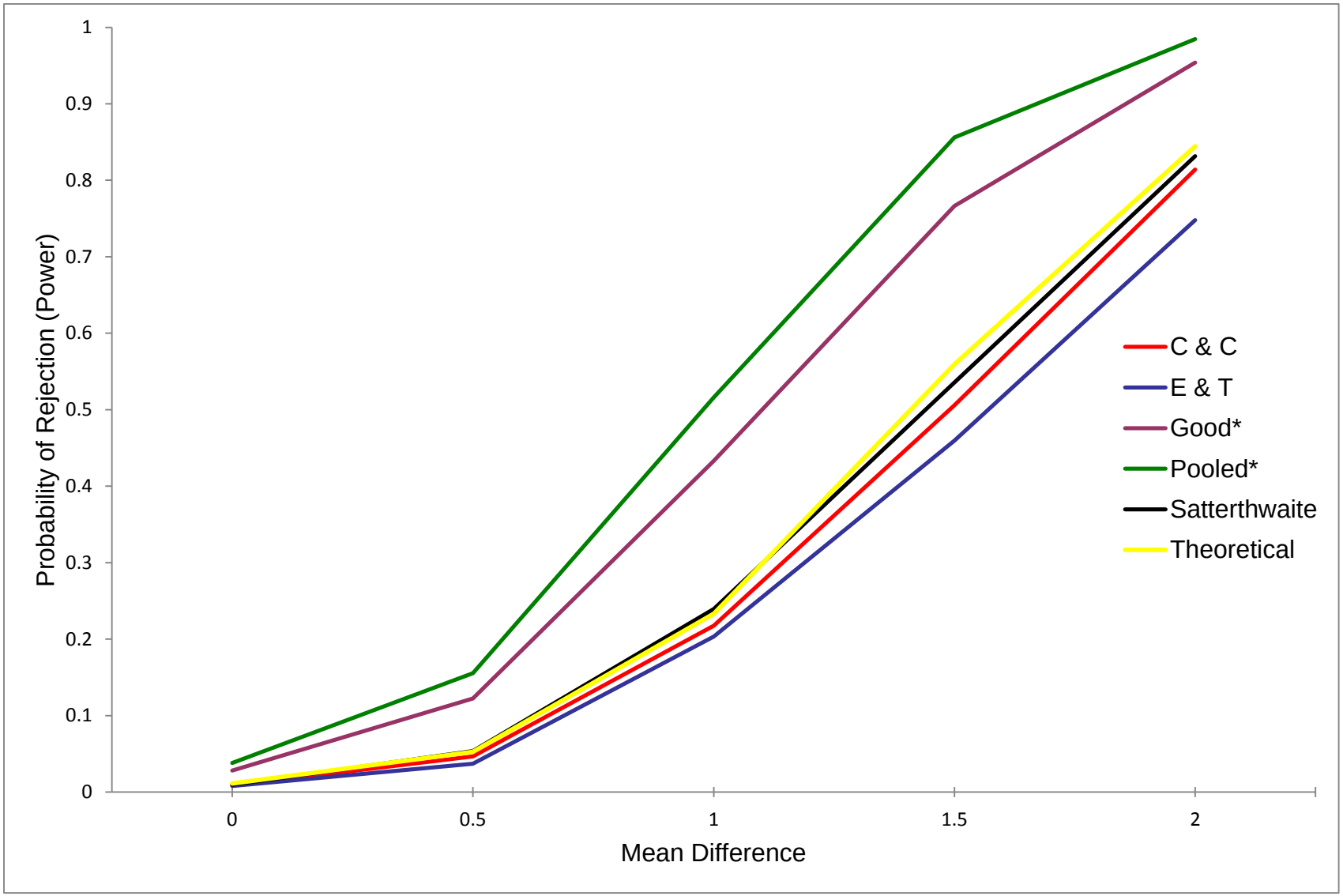

Figure A52. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

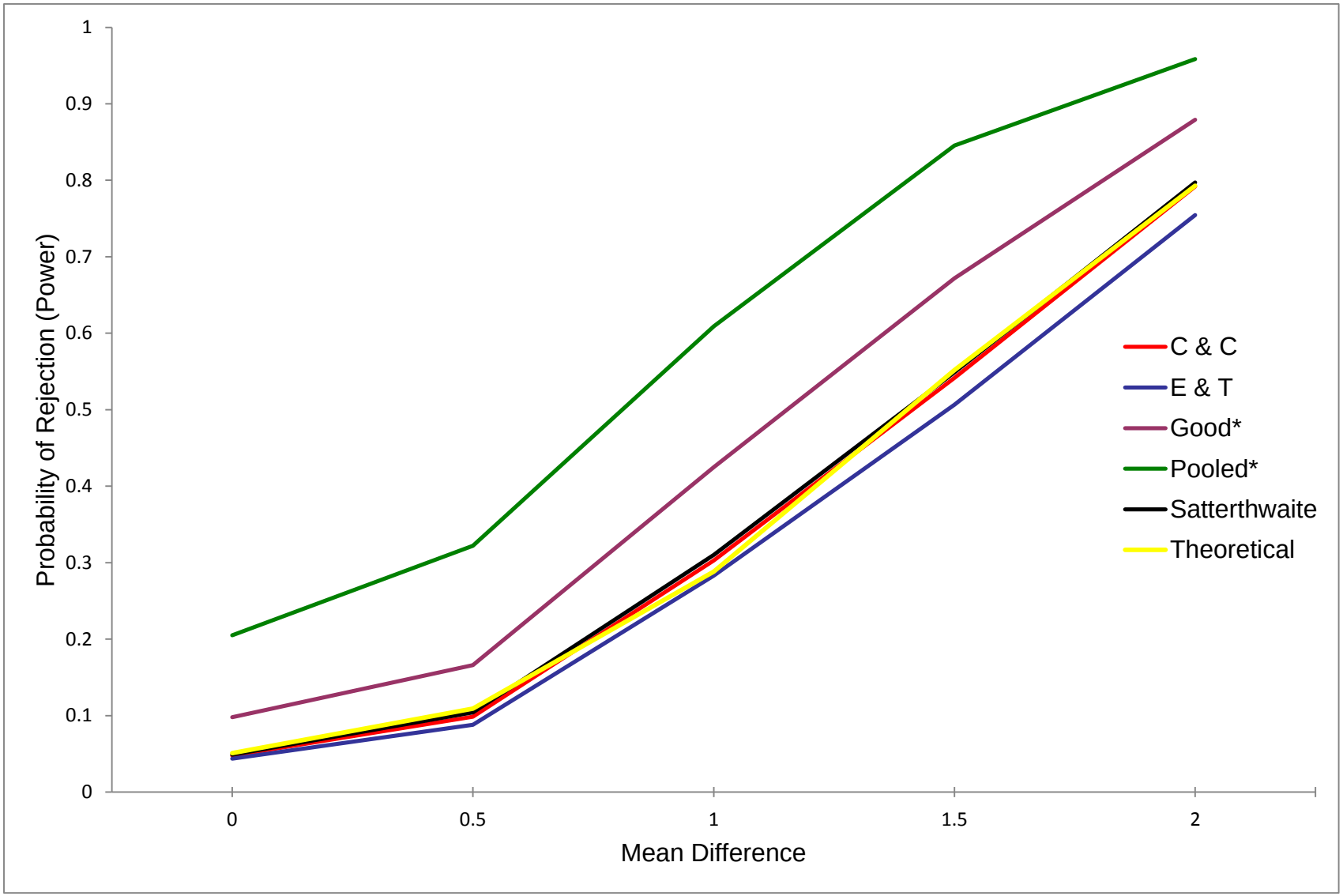

Figure A53. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

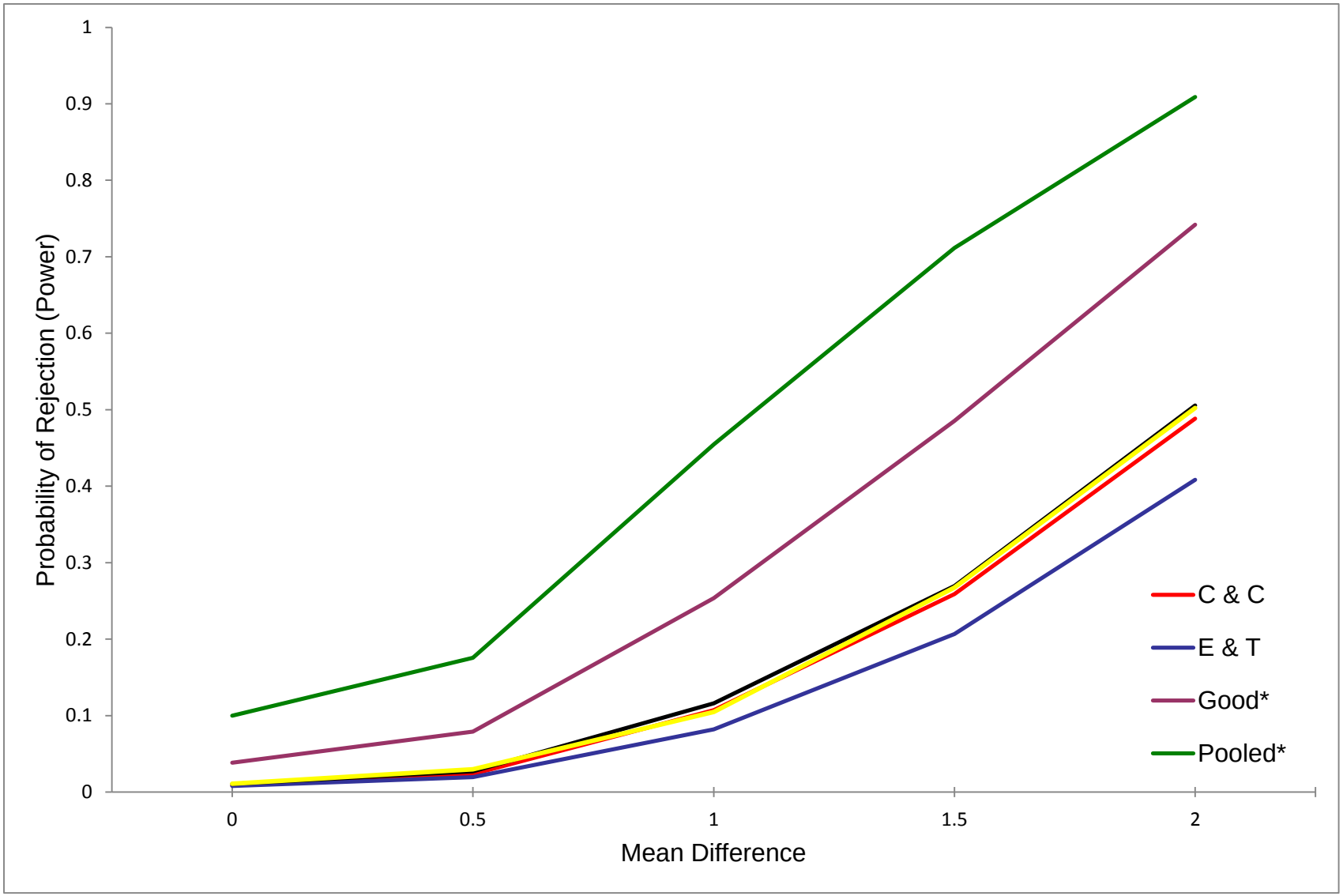

Figure A54. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

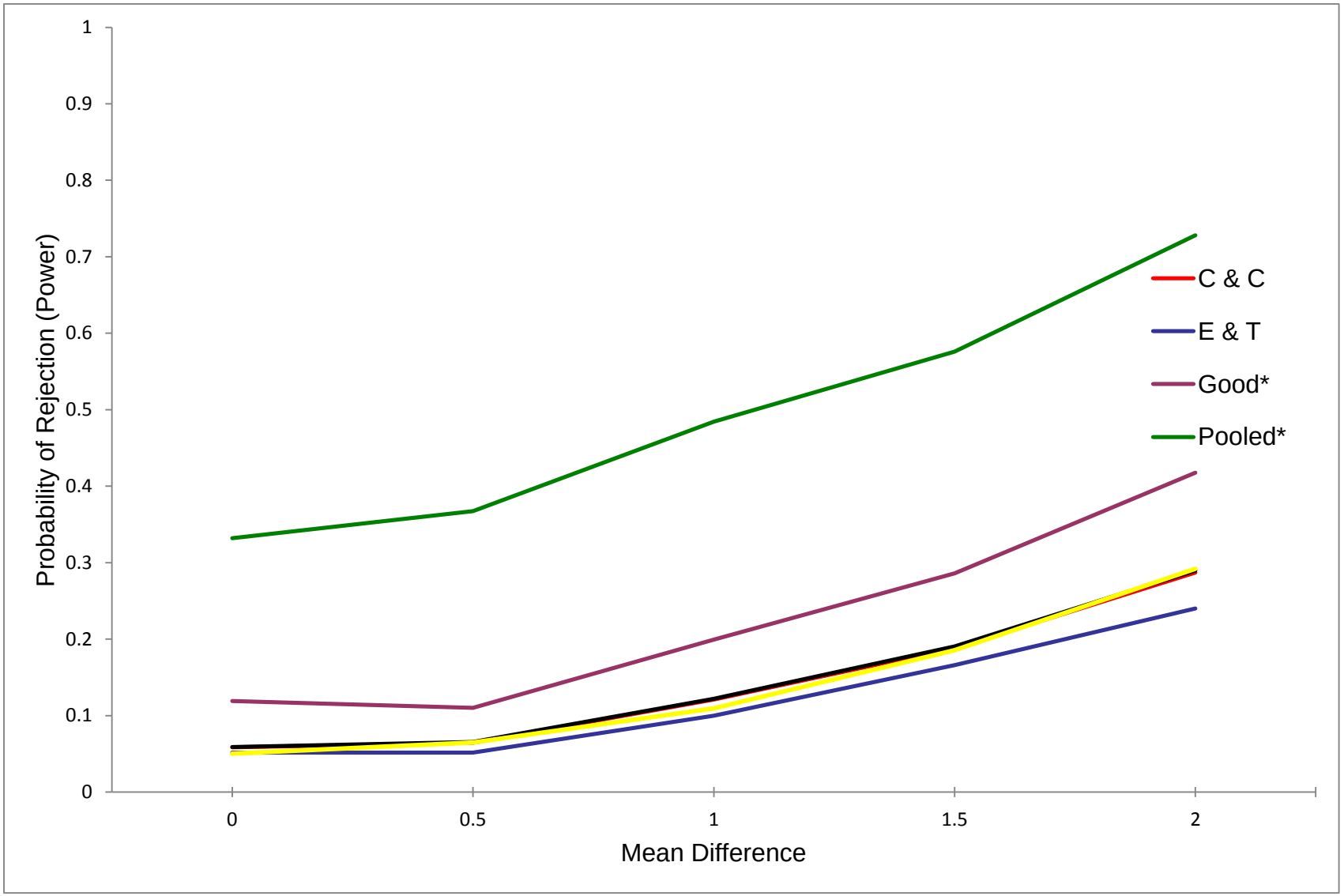

Figure A55. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

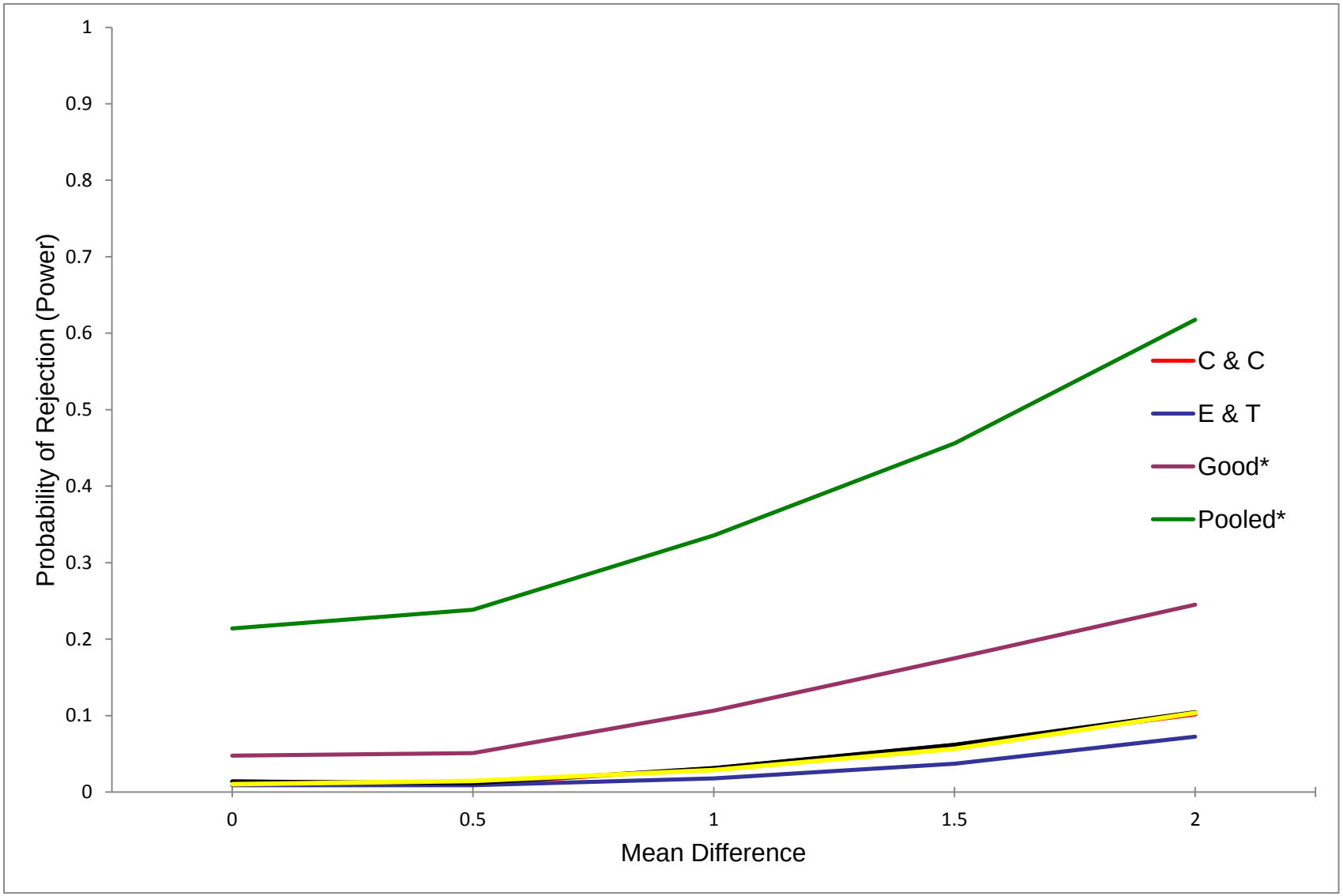

Figure A56. Power curves for unequal group sample sizes when $n_1 = 10$, $n_2 = 50$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

APPENDIX B: TYPE I ERROR RATE TABLES, POWER TABLES, AND POWER CURVES, WHEN THE SAMPLE SIZE OF GROUP 1 (n_1) WAS 25

Sample-size Ratio was 1.0 (i.e., Equal Sample Size or $n_1 = n_2 = 25$)

Table B1

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0455	0.0055
E & T	0.0455	0.0045
Good	0.0670*	0.0170
Pooled	0.0500	0.0085
Satterthwaite	0.0480	0.0060

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B2

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0380	0.0080
E & T	0.0400	0.0090
Good	0.0495	0.0155
Pooled	0.0415	0.0100
Satterthwaite	0.0405	0.0085

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B3

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0380	0.0065
E & T	0.0415	0.0075
Good	0.0505	0.0100
Pooled	0.0425	0.0080
Satterthwaite	0.0425	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B4

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0460	0.0070
E & T	0.0515	0.0095
Good	0.0635	0.0120
Pooled	0.0515	0.0085
Satterthwaite	0.0515	0.0085

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B5

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0455	0.0100
E & T	0.0500	0.0115
Good	0.0620	0.0170
Pooled	0.0510	0.0125
Satterthwaite	0.0505	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B6

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0545	0.0100
E & T	0.0565	0.0115
Good	0.0735*	0.0220*
Pooled	0.0620	0.0125
Satterthwaite	0.0585	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B7

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0420	0.0060
E & T	0.0415	0.0040
Good	0.0595	0.0155
Pooled	0.0465	0.0085
Satterthwaite	0.0430	0.0065

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B8

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0455	0.0380	0.0380	0.0460	0.0455	0.0545	0.0420
	0.5	0.6440	0.5870	0.5050	0.3890	0.2935	0.1835	0.0865
	1.0	0.9940	0.9940	0.9765	0.9195	0.7805	0.5795	0.2065
	1.5	1.0000	1.0000	1.0000	0.9995	0.9890	0.8950	0.4255
	2.0	1.0000	1.0000	1.0000	1.0000	0.9990	0.9935	0.6430
Efron & Tibshirani	0.0	0.0455	0.0400	0.0415	0.0515	0.0500	0.0565	0.0415
	0.5	0.6395	0.5940	0.5230	0.4095	0.3040	0.1895	0.0865
	1.0	0.9935	0.9945	0.9780	0.9255	0.7940	0.5880	0.2050
	1.5	1.0000	1.0000	1.0000	1.0000	0.9900	0.8985	0.4195
	2.0	1.0000	1.0000	1.0000	1.0000	0.9995	0.9935	0.6425
Good	0.0	0.0670	0.0495	0.0505	0.0635	0.0620	0.0735	0.0595
	0.5	0.6895	0.6365	0.5665	0.4525	0.3480	0.2195	0.1165
	1.0	0.9955	0.9960	0.9850	0.9350	0.8240	0.6275	0.2475
	1.5	1.0000	1.0000	1.0000	1.0000	0.9930	0.9155	0.4720
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9950	0.6955
Pooled	0.0	0.0500	0.0415	0.0425	0.0515	0.0510	0.0620	0.0465
	0.5	0.6605	0.6045	0.5290	0.4130	0.3075	0.1965	0.0975
	1.0	0.9940	0.9940	0.9805	0.9265	0.7960	0.5970	0.2205
	1.5	1.0000	1.0000	1.0000	1.0000	0.9915	0.9050	0.4410
	2.0	1.0000	1.0000	1.0000	1.0000	0.9995	0.9935	0.6630
Satterthwaite	0.0	0.0480	0.0405	0.0425	0.0515	0.0505	0.0585	0.0430
	0.5	0.6475	0.5965	0.5265	0.4120	0.3045	0.1900	0.0870
	1.0	0.9940	0.9940	0.9800	0.9265	0.7940	0.5885	0.2090
	1.5	1.0000	1.0000	1.0000	1.0000	0.9915	0.9020	0.4285
	2.0	1.0000	1.0000	1.0000	1.0000	0.9995	0.9935	0.6465

Table B9

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 25$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0055	0.0080	0.0065	0.0070	0.0100	0.0100	0.0060
	0.5	0.3760	0.3065	0.2550	0.1595	0.1030	0.0515	0.0265
	1.0	0.9710	0.9465	0.9040	0.7655	0.5195	0.3010	0.0690
	1.5	1.0000	1.0000	1.0000	0.9935	0.9330	0.6980	0.1910
	2.0	1.0000	1.0000	1.0000	1.0000	0.9965	0.9475	0.3720
Efron & Tibshirani	0.0	0.0045	0.0090	0.0075	0.0095	0.0115	0.0115	0.0040
	0.5	0.3675	0.3185	0.2765	0.1800	0.1160	0.0550	0.0265
	1.0	0.9660	0.9450	0.9115	0.7870	0.5465	0.3110	0.0690
	1.5	1.0000	1.0000	1.0000	0.9940	0.9400	0.7060	0.1875
	2.0	1.0000	1.0000	1.0000	1.0000	0.9965	0.9550	0.3590
Good	0.0	0.0170	0.0155	0.0100	0.0120	0.0170	0.0220	0.0155
	0.5	0.4725	0.4100	0.3445	0.2325	0.1575	0.0870	0.0410
	1.0	0.9845	0.9705	0.9430	0.8355	0.6160	0.3915	0.1060
	1.5	1.0000	1.0000	1.0000	0.9975	0.9595	0.7835	0.2700
	2.0	1.0000	1.0000	1.0000	1.0000	0.9970	0.9725	0.4740
Pooled	0.0	0.0085	0.0100	0.0080	0.0085	0.0125	0.0125	0.0085
	0.5	0.4135	0.3500	0.2920	0.1935	0.1235	0.0635	0.0315
	1.0	0.9770	0.9575	0.9195	0.7960	0.5610	0.3325	0.0825
	1.5	1.0000	1.0000	1.0000	0.9950	0.9445	0.7305	0.2270
	2.0	1.0000	1.0000	1.0000	1.0000	0.9970	0.9620	0.4155
Satterthwaite	0.0	0.0060	0.0085	0.0080	0.0085	0.0120	0.0120	0.0065
	0.5	0.3845	0.3360	0.2865	0.1905	0.1210	0.0575	0.0285
	1.0	0.9725	0.9525	0.9160	0.7935	0.5545	0.3200	0.0720
	1.5	1.0000	1.0000	1.0000	0.9950	0.9435	0.7205	0.2000
	2.0	1.0000	1.0000	1.0000	1.0000	0.9970	0.9580	0.3835

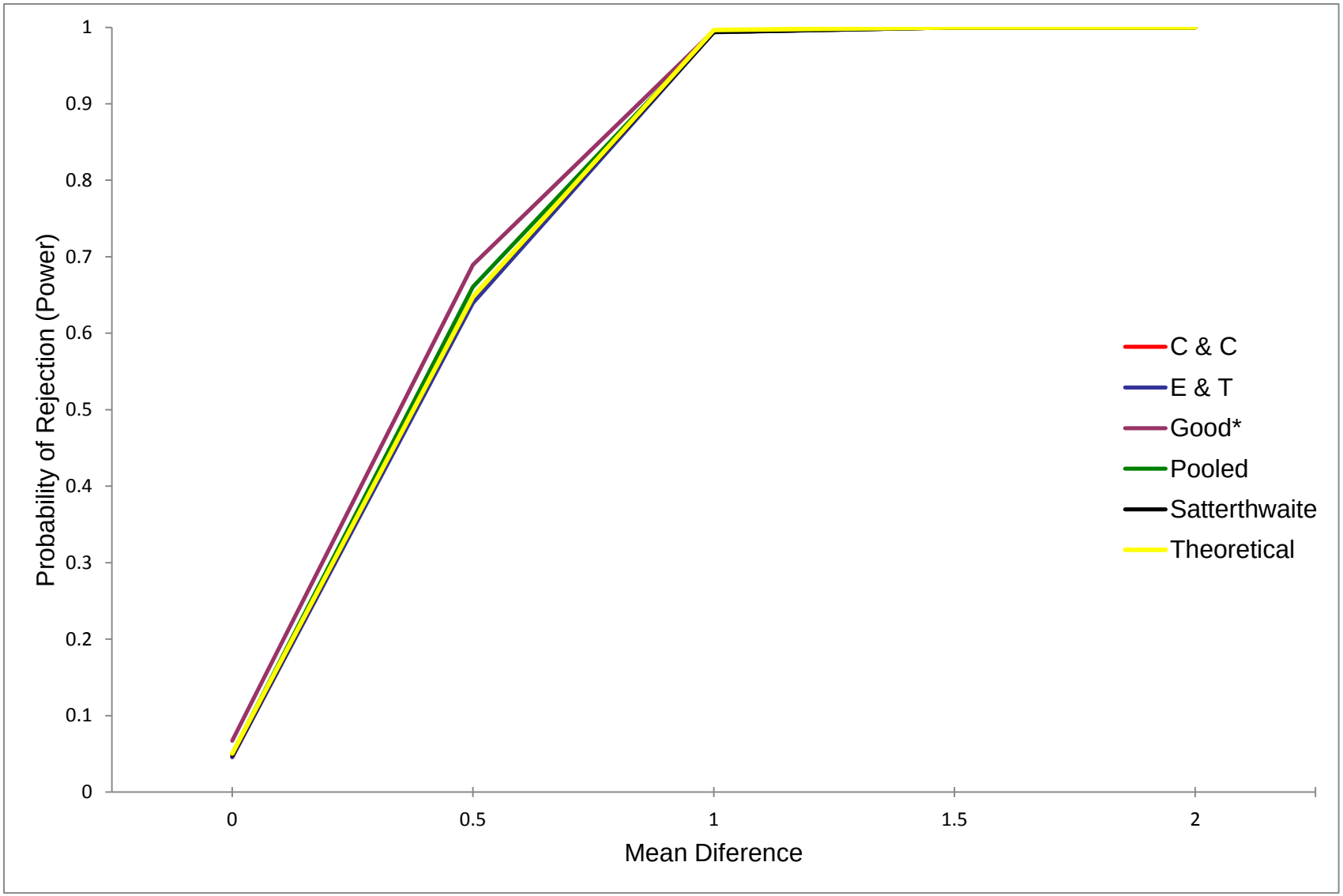

Figure B1. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

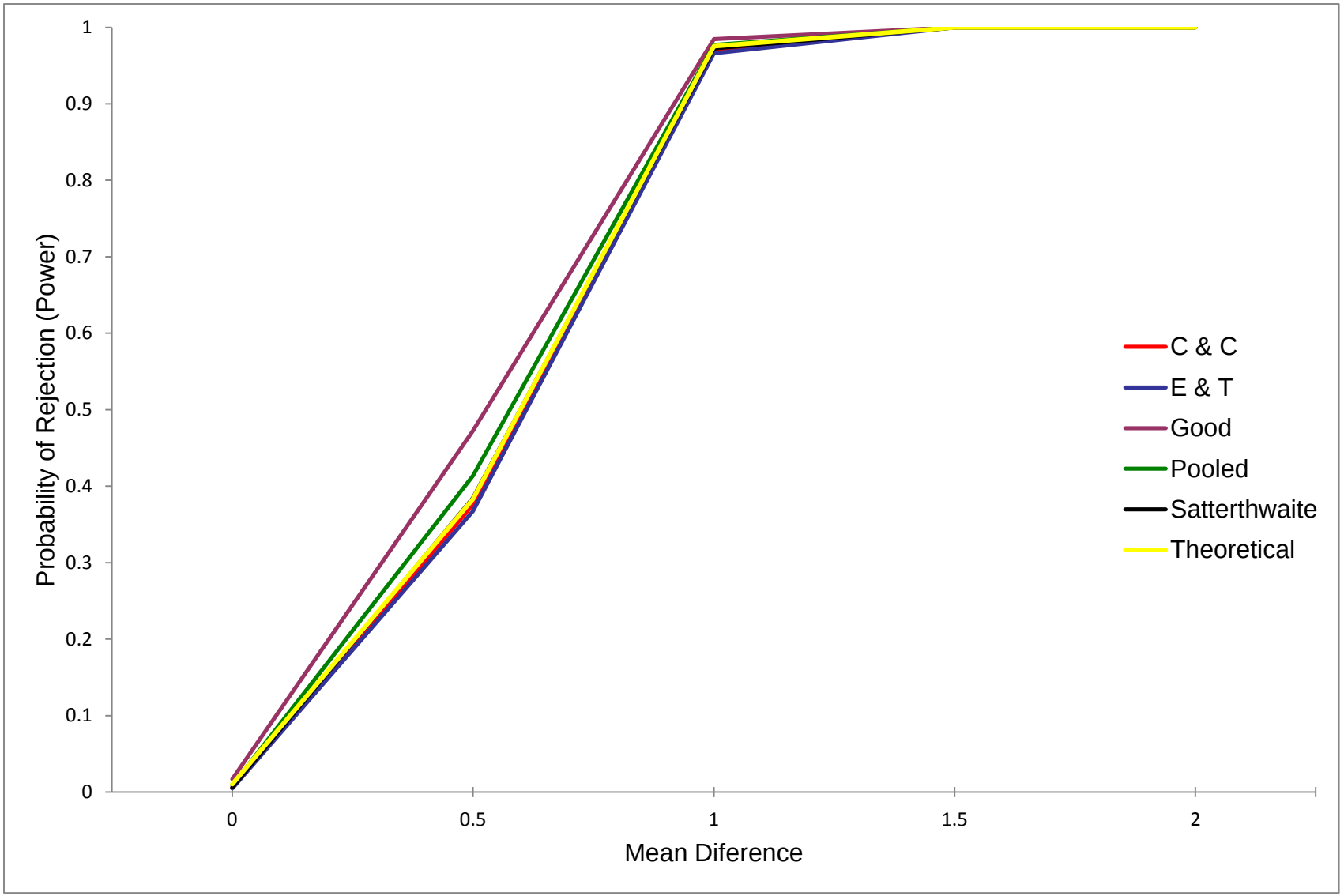

Figure B2. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

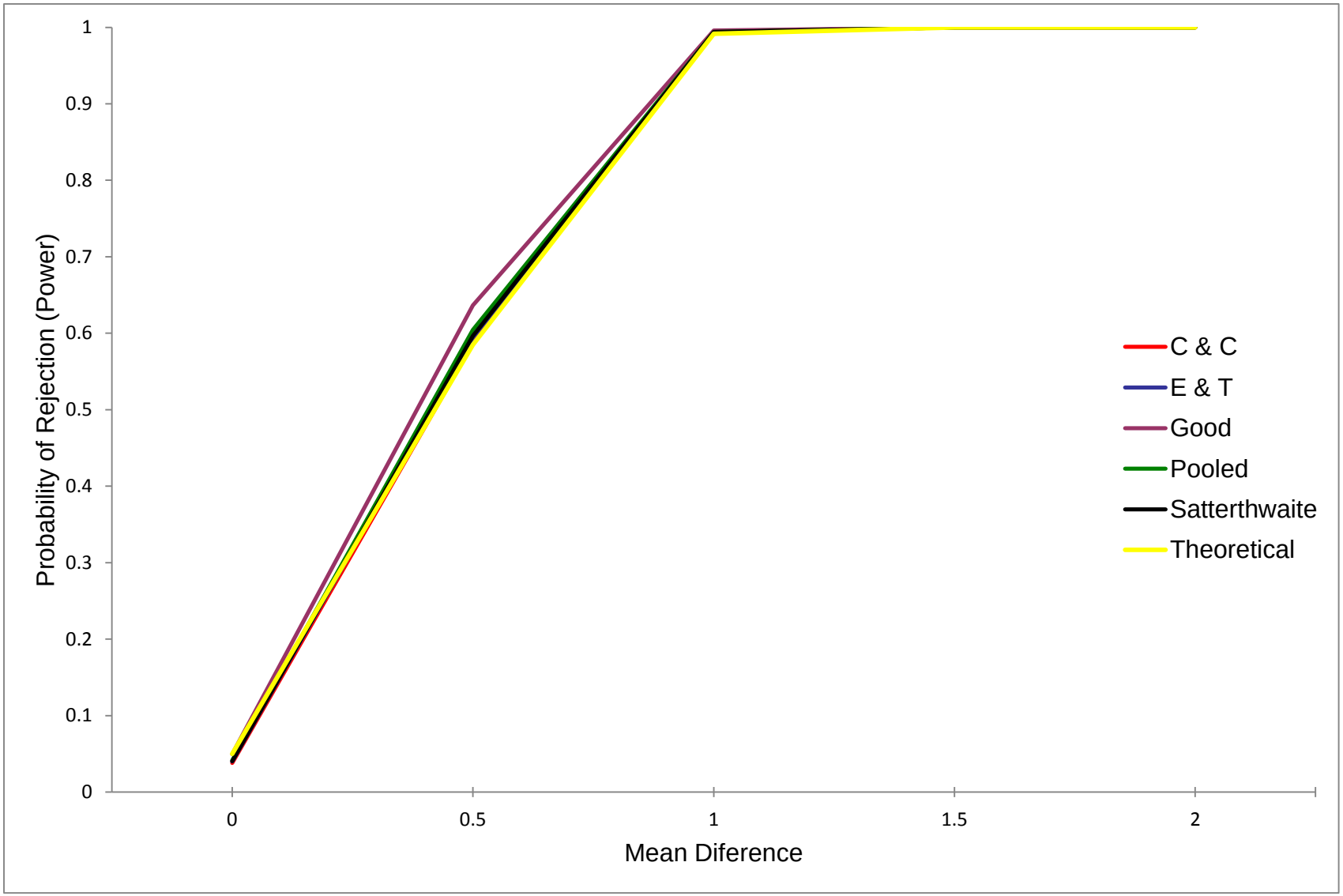

Figure B3. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

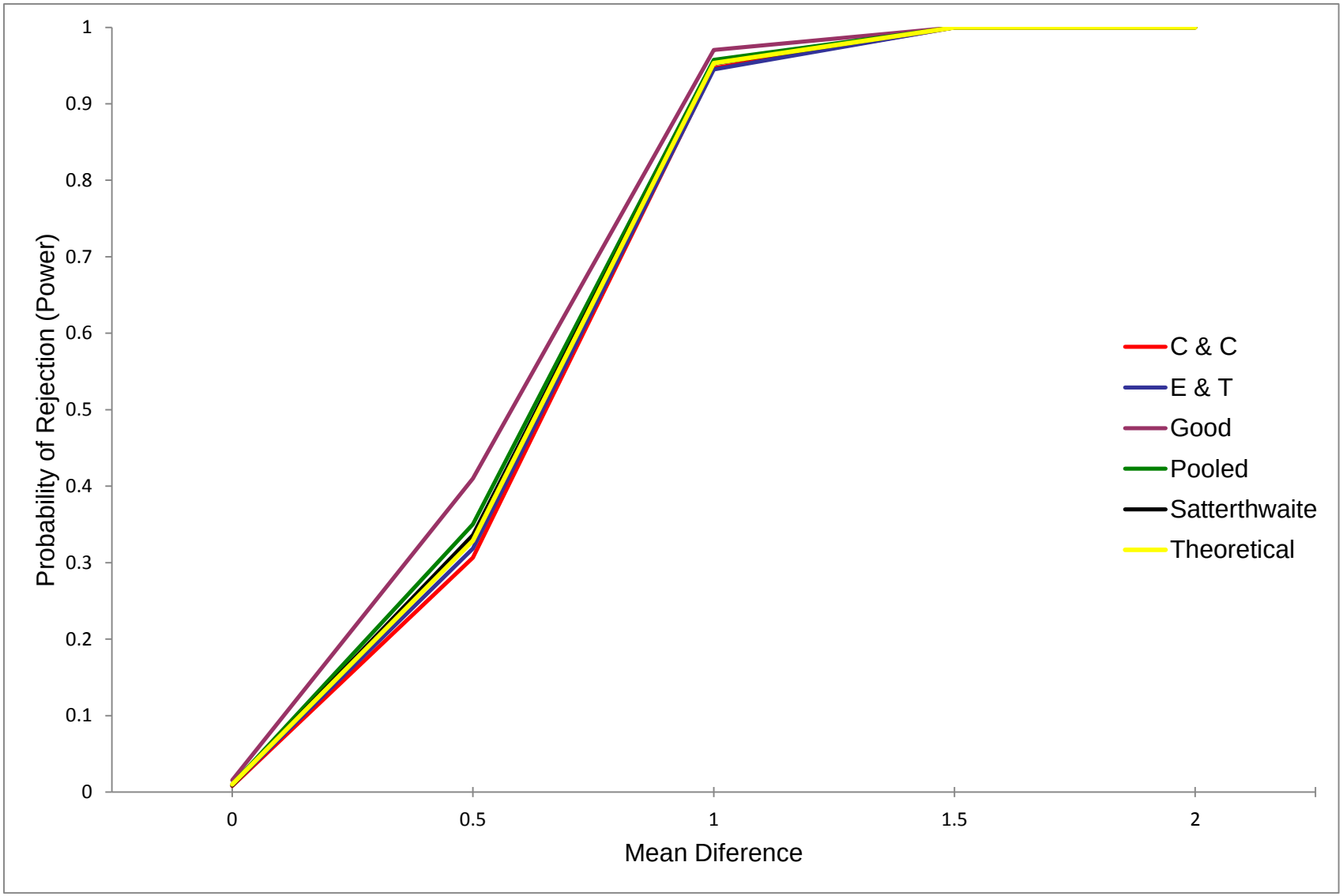

Figure B4. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

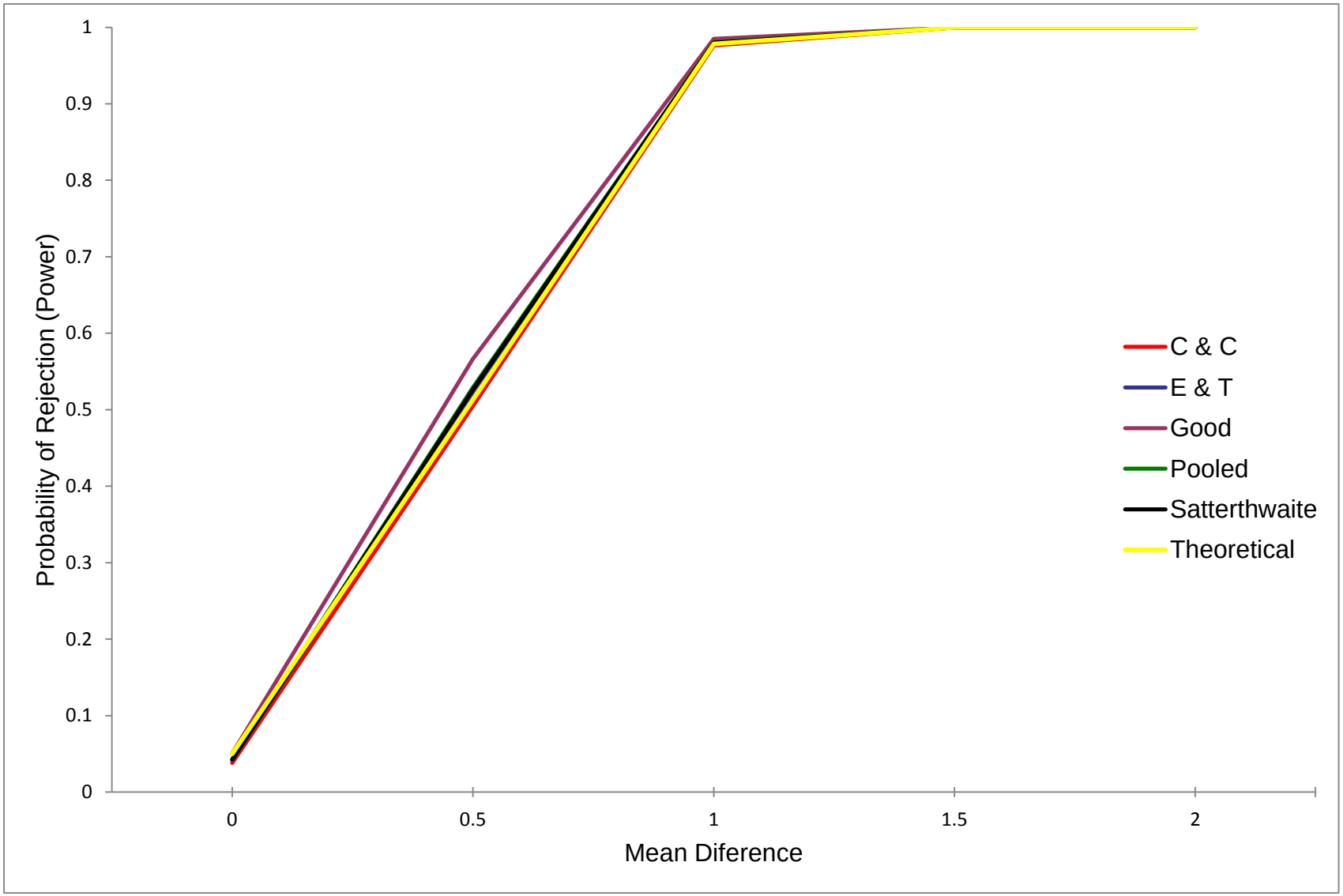

Figure B5. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

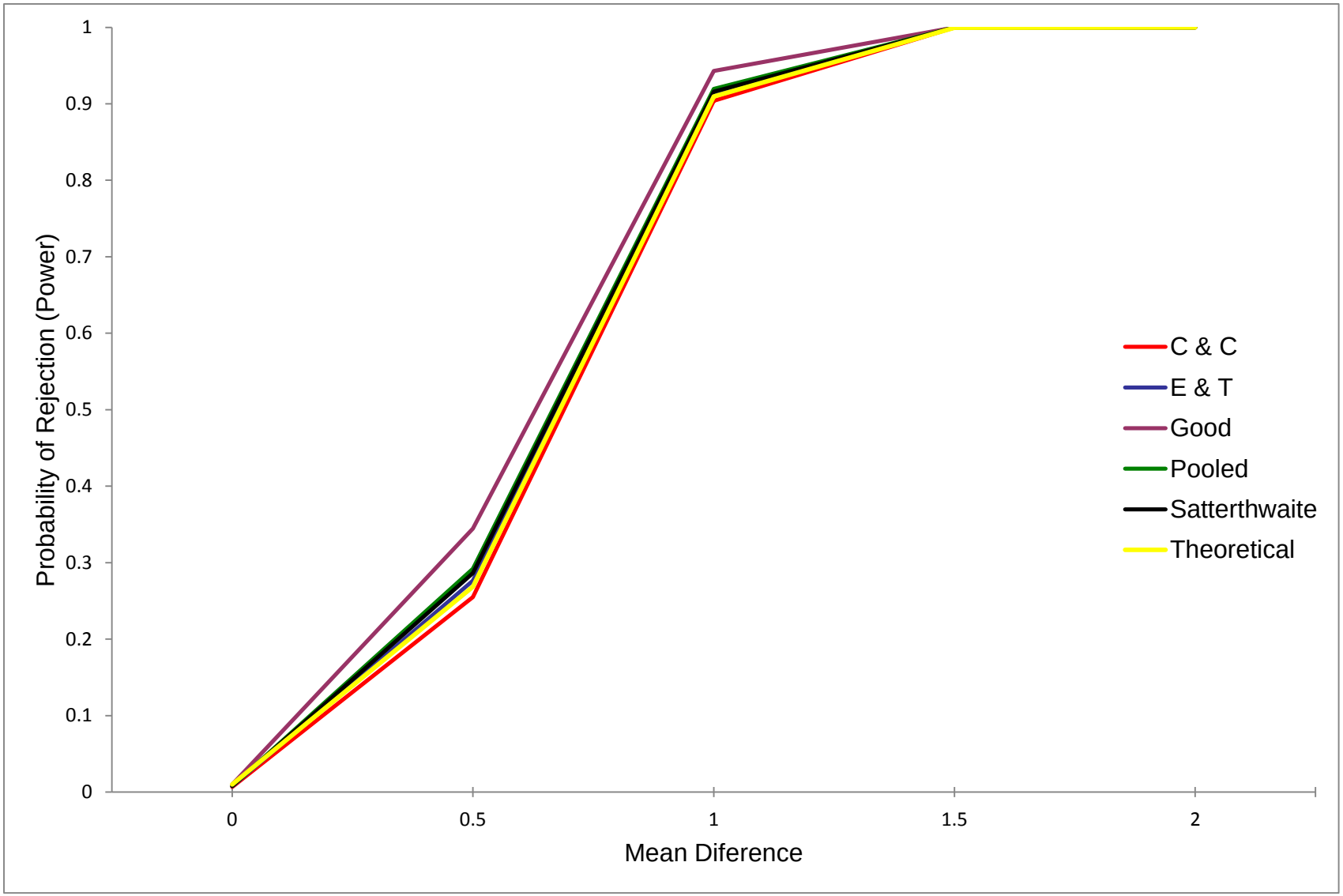

Figure B6. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

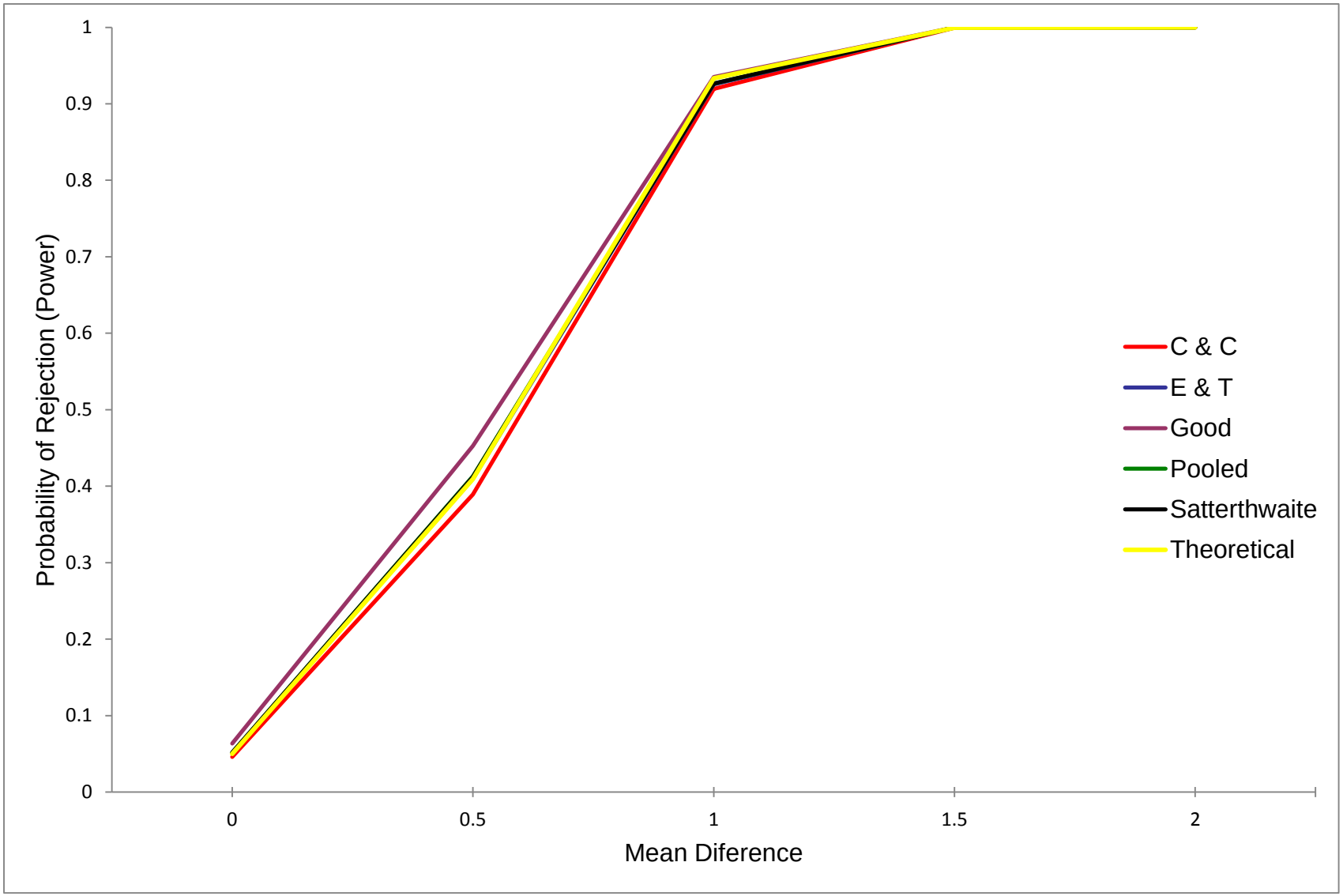

Figure B7. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

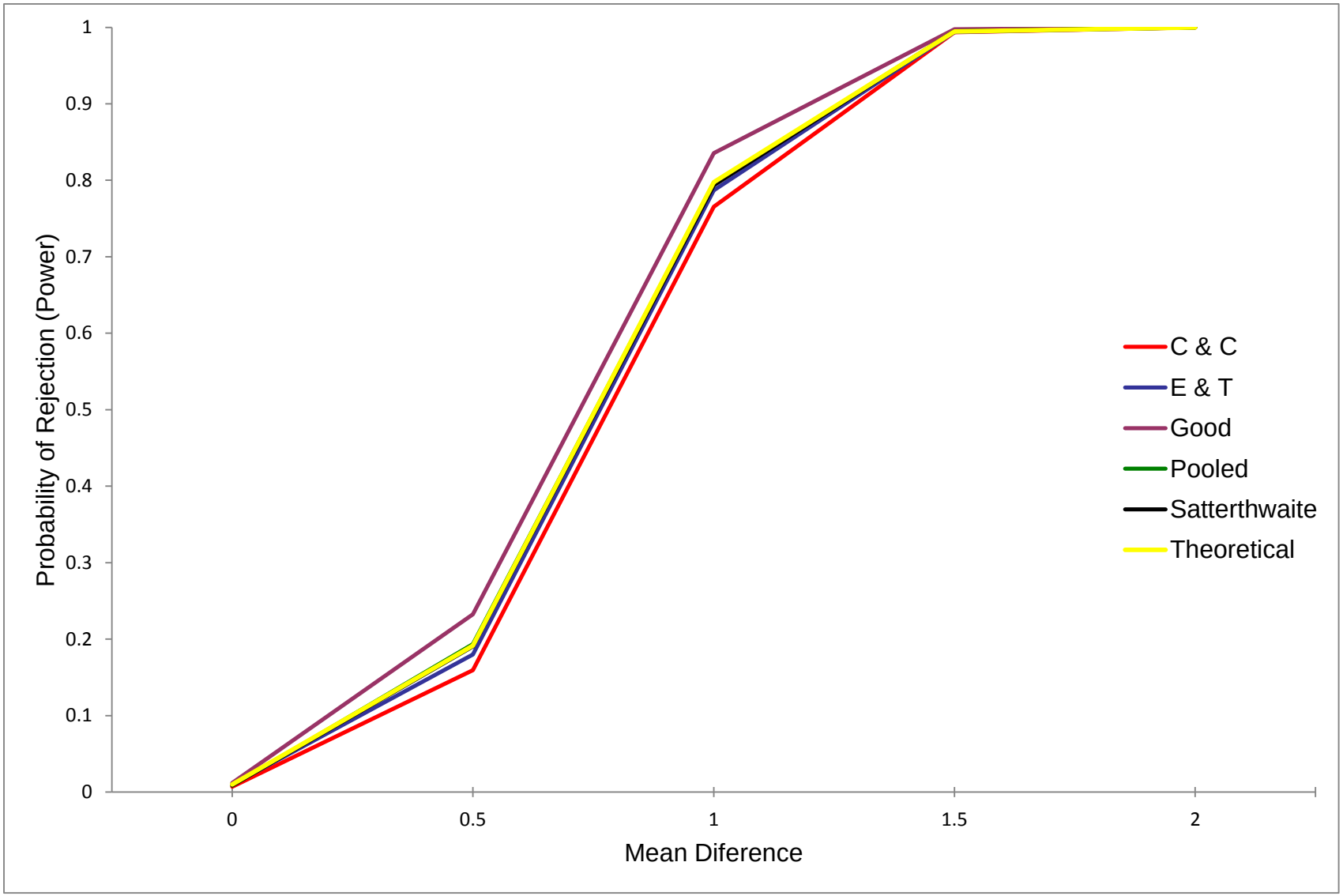

Figure B8. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

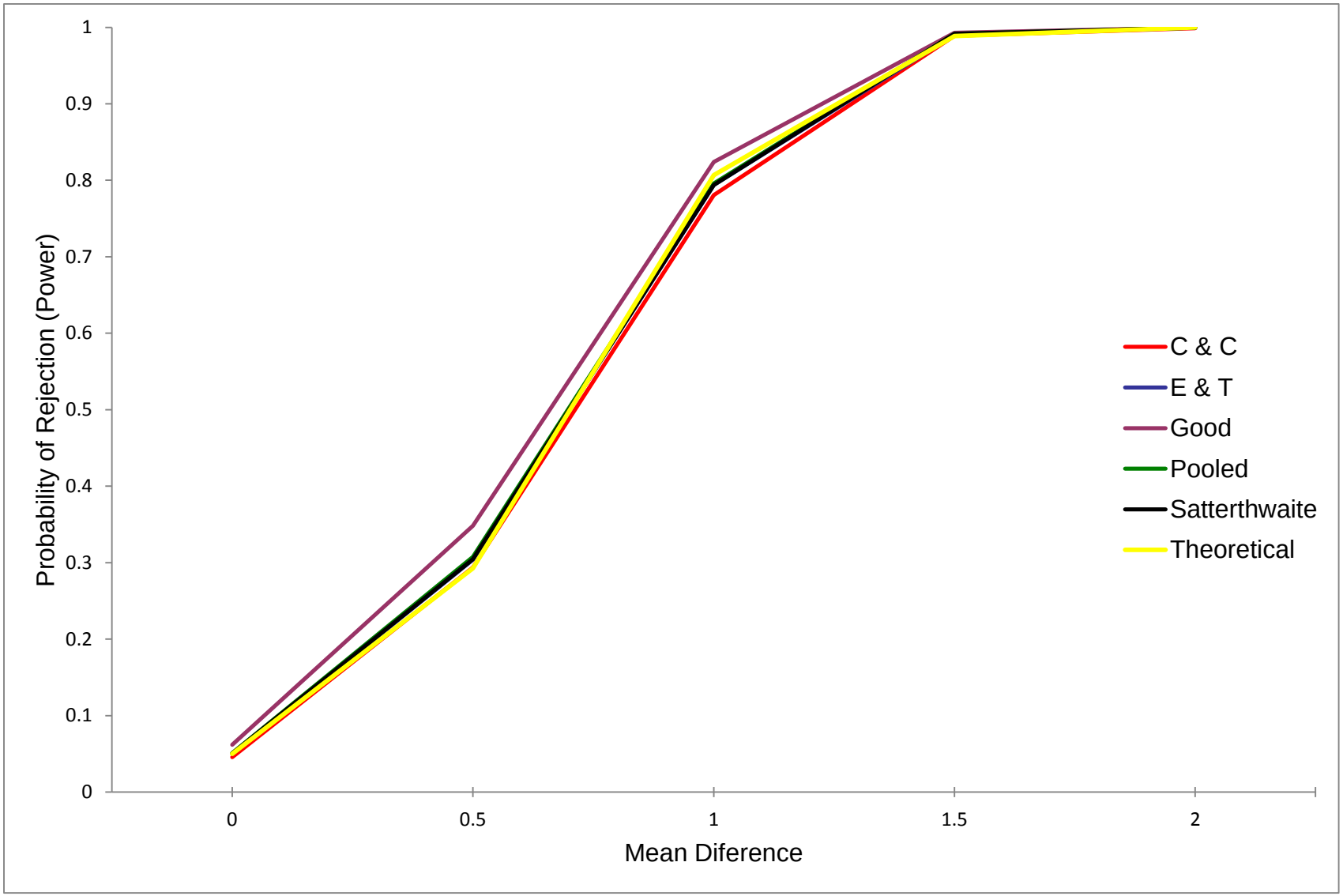

Figure B9. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

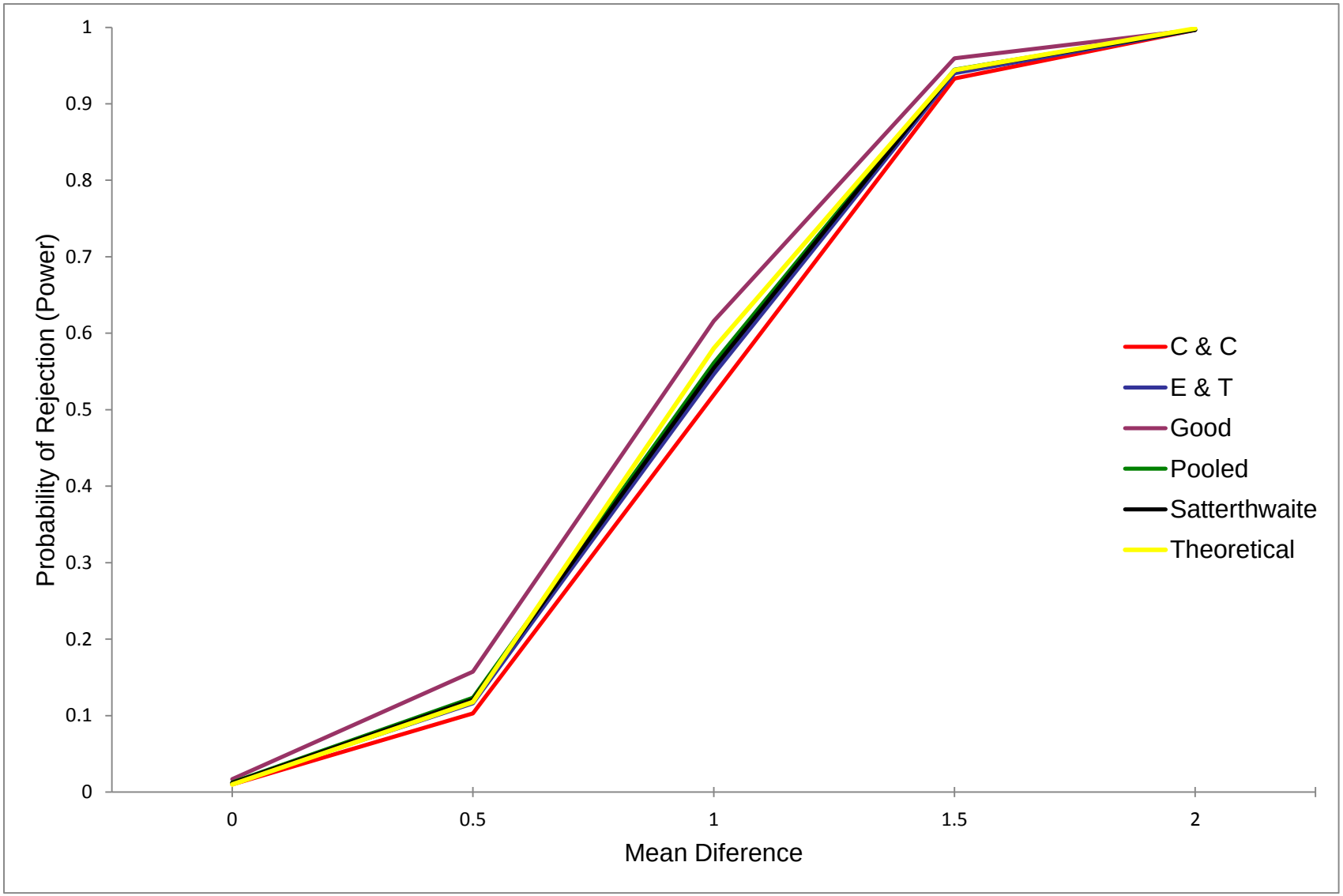

Figure B10 Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

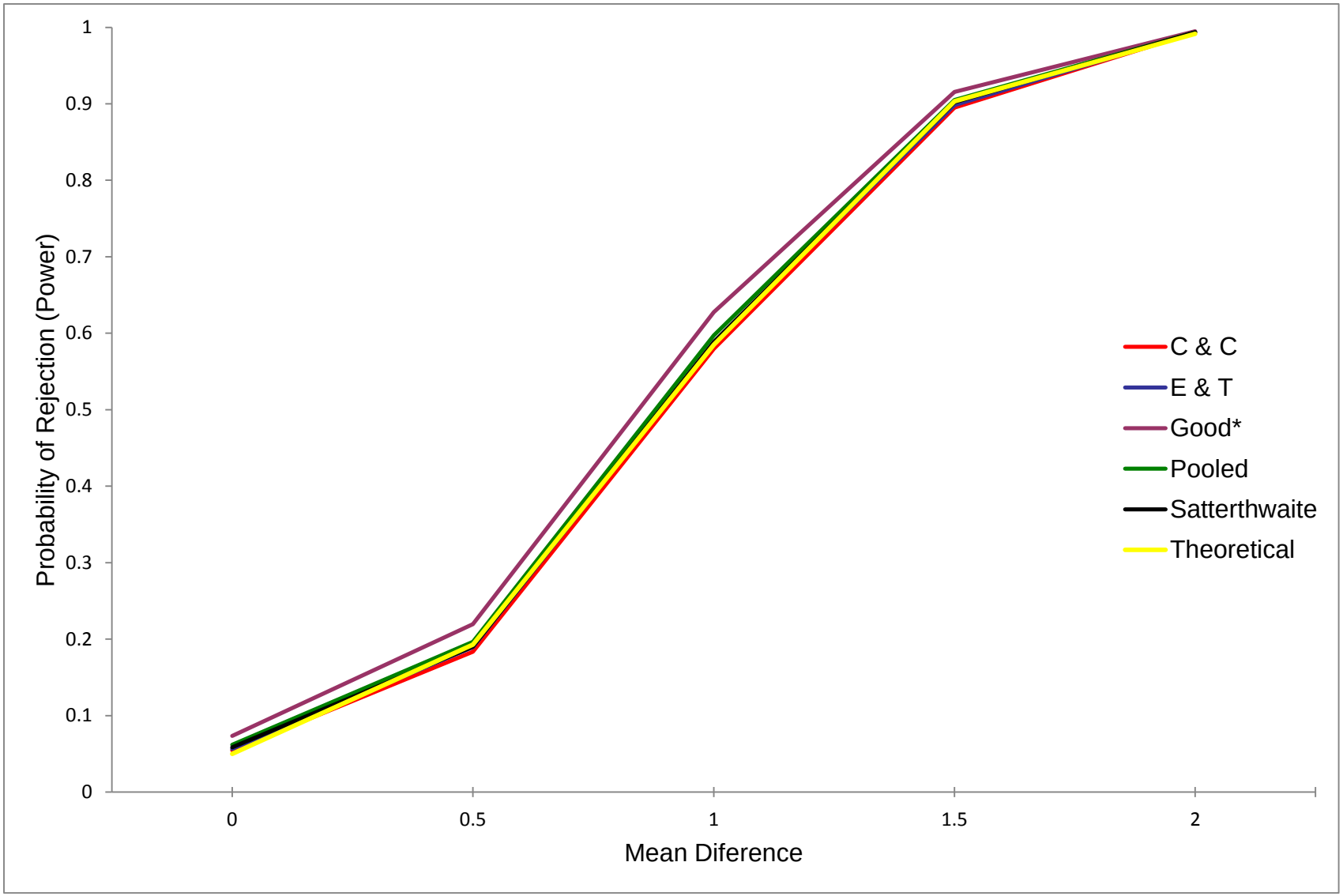

Figure B11. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

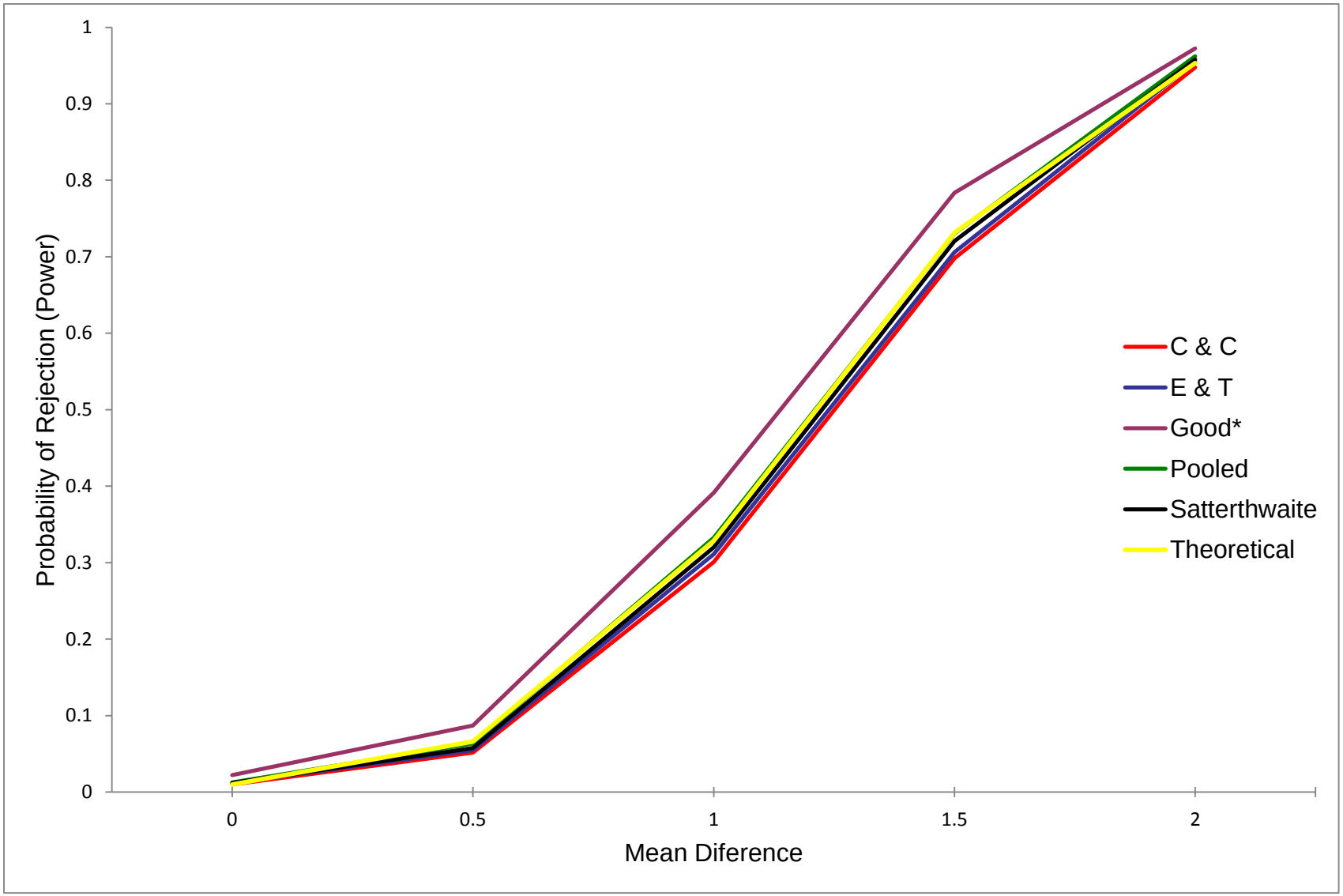

Figure B12. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

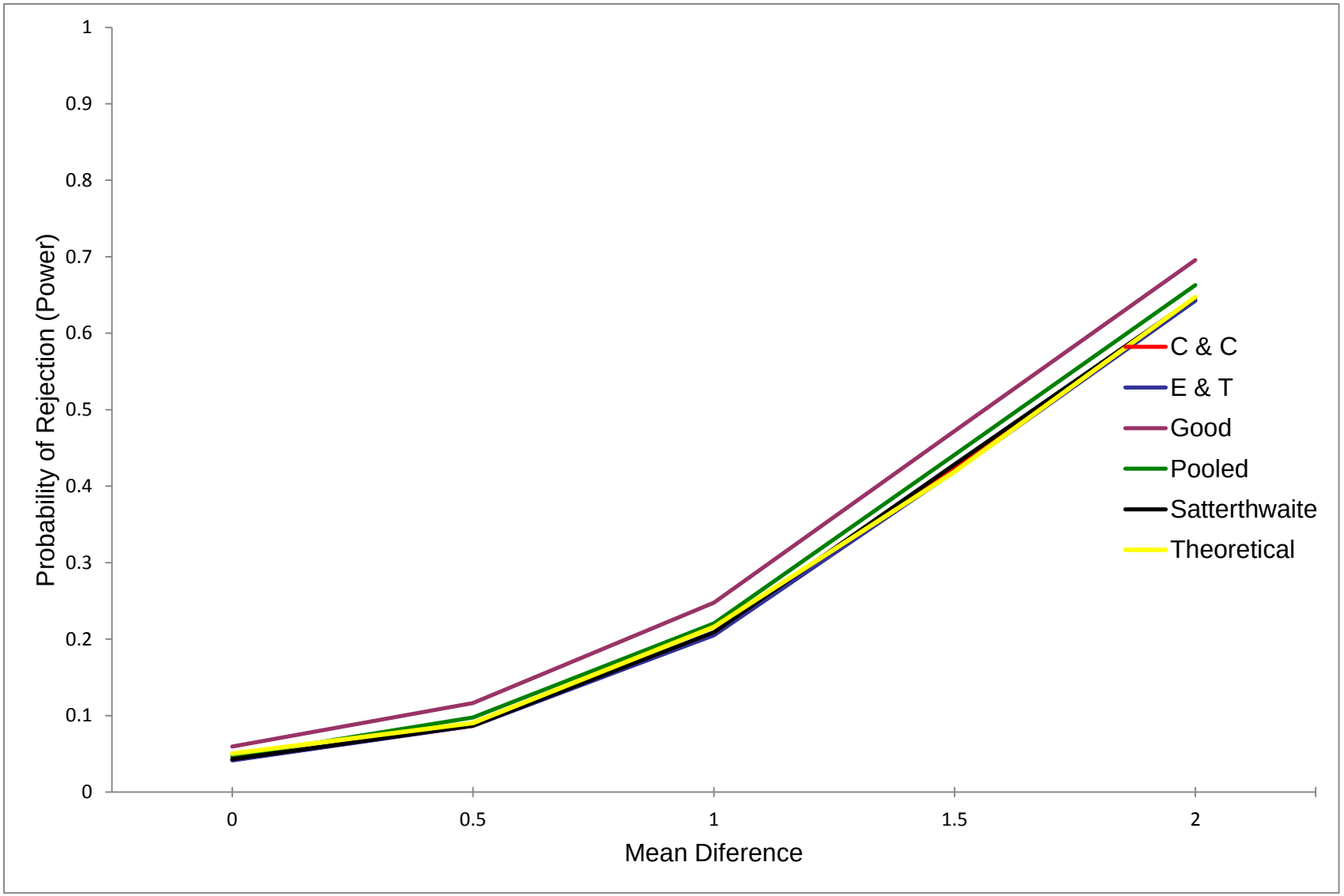

Figure B13. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

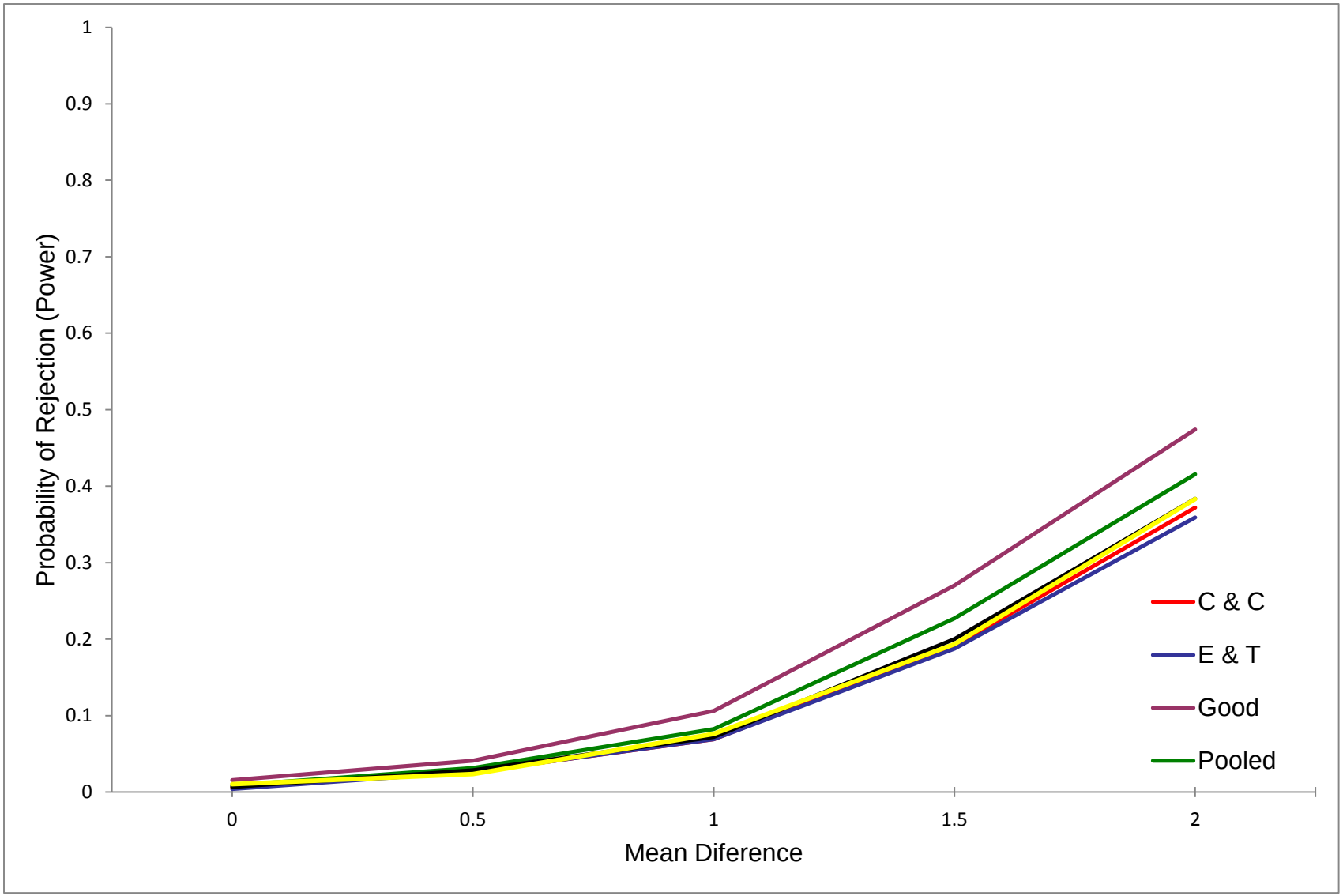

Figure B14. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 25$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 1.5 (i.e., $n_1 = 25$, $n_2 = 38$)

Table B10

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/16, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0590	0.0095
E & T	0.0600	0.0105
Good	0.0925*	0.0350*
Pooled	0.0355	0.0045
Satterthwaite	0.0610	0.0135

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B11

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0420	0.0060
E & T	0.0505	0.0090
Good	0.0745*	0.0225*
Pooled	0.0320*	0.0050
Satterthwaite	0.0530	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B12

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0450	0.0045
E & T	0.0545	0.0080
Good	0.0820*	0.0275*
Pooled	0.0460	0.0065
Satterthwaite	0.0565	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B13

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0380	0.0075
E & T	0.0435	0.0085
Good	0.0720*	0.0230*
Pooled	0.0465	0.0110
Satterthwaite	0.0440	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B14

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0395	0.0050
E & T	0.0440	0.0060
Good	0.0765*	0.0275*
Pooled	0.0640	0.0170
Satterthwaite	0.0470	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B15

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0365	0.0035
E & T	0.0355	0.0045
Good	0.0770*	0.0290*
Pooled	0.0760*	0.0205*
Satterthwaite	0.0410	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B16

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0505	0.0080
E & T	0.0425	0.0075
Good	0.0960*	0.0405*
Pooled	0.1140*	0.0395*
Satterthwaite	0.0525	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B17

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0475	0.0410	0.0465	0.0495	0.0535	0.0530	0.0535
	0.5	0.8240	0.7180	0.6110	0.4585	0.2955	0.1890	0.0870
	1.0	1.0000	0.9990	0.9940	0.9695	0.8430	0.5985	0.2365
	1.5	1.0000	1.0000	1.0000	1.0000	0.9910	0.9170	0.4230
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9885	0.6540
Efron & Tibshirani	0.0	0.0490	0.0460	0.0540	0.0540	0.0560	0.0530	0.0530
	0.5	0.8270	0.7265	0.6275	0.4730	0.3055	0.1905	0.0885
	1.0	1.0000	1.0000	0.9945	0.9720	0.8455	0.6000	0.2330
	1.5	1.0000	1.0000	1.0000	1.0000	0.9915	0.9170	0.4190
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9885	0.6515
Good	0.0	0.0620	0.0550	0.0635	0.0640	0.0680	0.0675	0.0760
	0.5	0.8470	0.7480	0.6625	0.5020	0.3390	0.2300	0.1185
	1.0	1.0000	1.0000	0.9945	0.9805	0.8685	0.6405	0.2725
	1.5	1.0000	1.0000	1.0000	1.0000	0.9950	0.9410	0.4795
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9930	0.6980
Pooled	0.0	0.0195	0.0250	0.0385	0.0550	0.0760	0.0855	0.1190
	0.5	0.7060	0.6390	0.5775	0.4740	0.3630	0.2845	0.1755
	1.0	0.9985	0.9975	0.9910	0.9770	0.8795	0.7105	0.3555
	1.5	1.0000	1.0000	1.0000	1.0000	0.9960	0.9570	0.5880
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9950	0.7840
Satterthwaite	0.0	0.0500	0.0465	0.0535	0.0530	0.0565	0.0550	0.0535
	0.5	0.8260	0.7280	0.6295	0.4745	0.3070	0.1945	0.0890
	1.0	1.0000	0.9995	0.9940	0.9725	0.8485	0.6050	0.2395
	1.5	1.0000	1.0000	1.0000	1.0000	0.9915	0.9190	0.4260
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9890	0.6555

Table B18

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 38$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0085	0.0075	0.0085	0.0110	0.0110	0.0095	0.0120
	0.5	0.5975	0.4530	0.3450	0.2240	0.1175	0.0685	0.0275
	1.0	0.9975	0.9910	0.9650	0.8625	0.6290	0.3325	0.0835
	1.5	1.0000	1.0000	1.0000	0.9985	0.9600	0.7510	0.1960
	2.0	1.0000	1.0000	1.0000	1.0000	0.9995	0.9605	0.3930
Efron & Tibshirani	0.0	0.0090	0.0080	0.0105	0.0125	0.0110	0.0105	0.0120
	0.5	0.5980	0.4775	0.3710	0.2460	0.1285	0.0695	0.0250
	1.0	0.9965	0.9925	0.9730	0.8760	0.6345	0.3355	0.0790
	1.5	1.0000	1.0000	1.0000	0.9990	0.9620	0.7505	0.1800
	2.0	1.0000	1.0000	1.0000	1.0000	0.9995	0.9605	0.3735
Good	0.0	0.0155	0.0120	0.0150	0.0195	0.0175	0.0190	0.0230
	0.5	0.6585	0.5225	0.4220	0.2945	0.1730	0.0975	0.0470
	1.0	0.9980	0.9955	0.9785	0.9115	0.7055	0.4145	0.1260
	1.5	1.0000	1.0000	1.0000	0.9990	0.9765	0.8180	0.2635
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9770	0.4905
Pooled	0.0	0.0005	0.0040	0.0065	0.0150	0.0195	0.0245	0.0400
	0.5	0.4205	0.3580	0.3085	0.2550	0.1770	0.1250	0.0715
	1.0	0.9930	0.9855	0.9605	0.8900	0.7230	0.4840	0.2000
	1.5	1.0000	1.0000	1.0000	0.9990	0.9805	0.8615	0.3735
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9825	0.5980
Satterthwaite	0.0	0.0095	0.0090	0.0105	0.0140	0.0110	0.0110	0.0120
	0.5	0.6085	0.4810	0.3740	0.2480	0.1305	0.0720	0.0280
	1.0	0.9975	0.9925	0.9730	0.8830	0.6495	0.3480	0.0860
	1.5	1.0000	1.0000	1.0000	0.9990	0.9675	0.7640	0.2000
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9635	0.3990

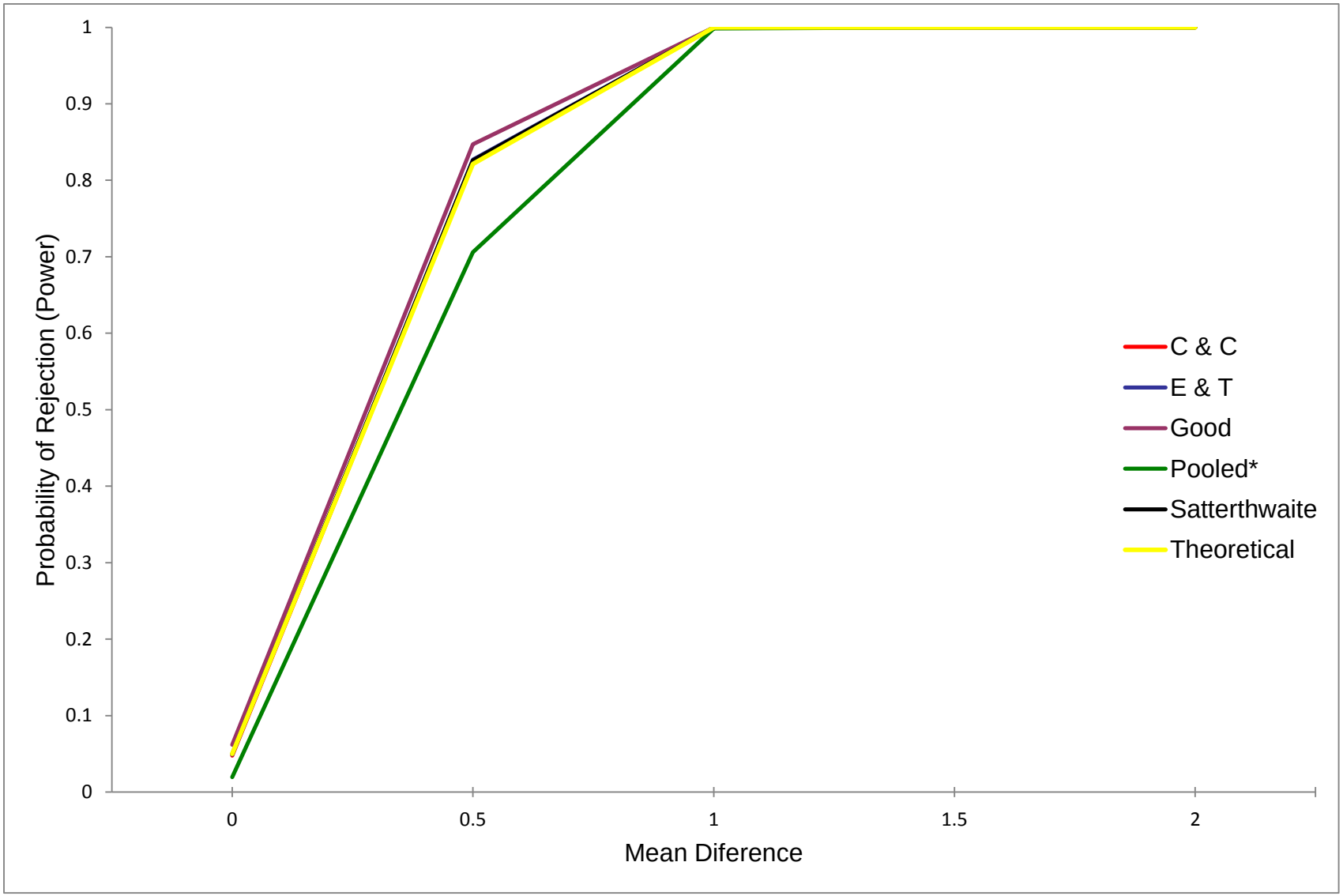

Figure B15. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

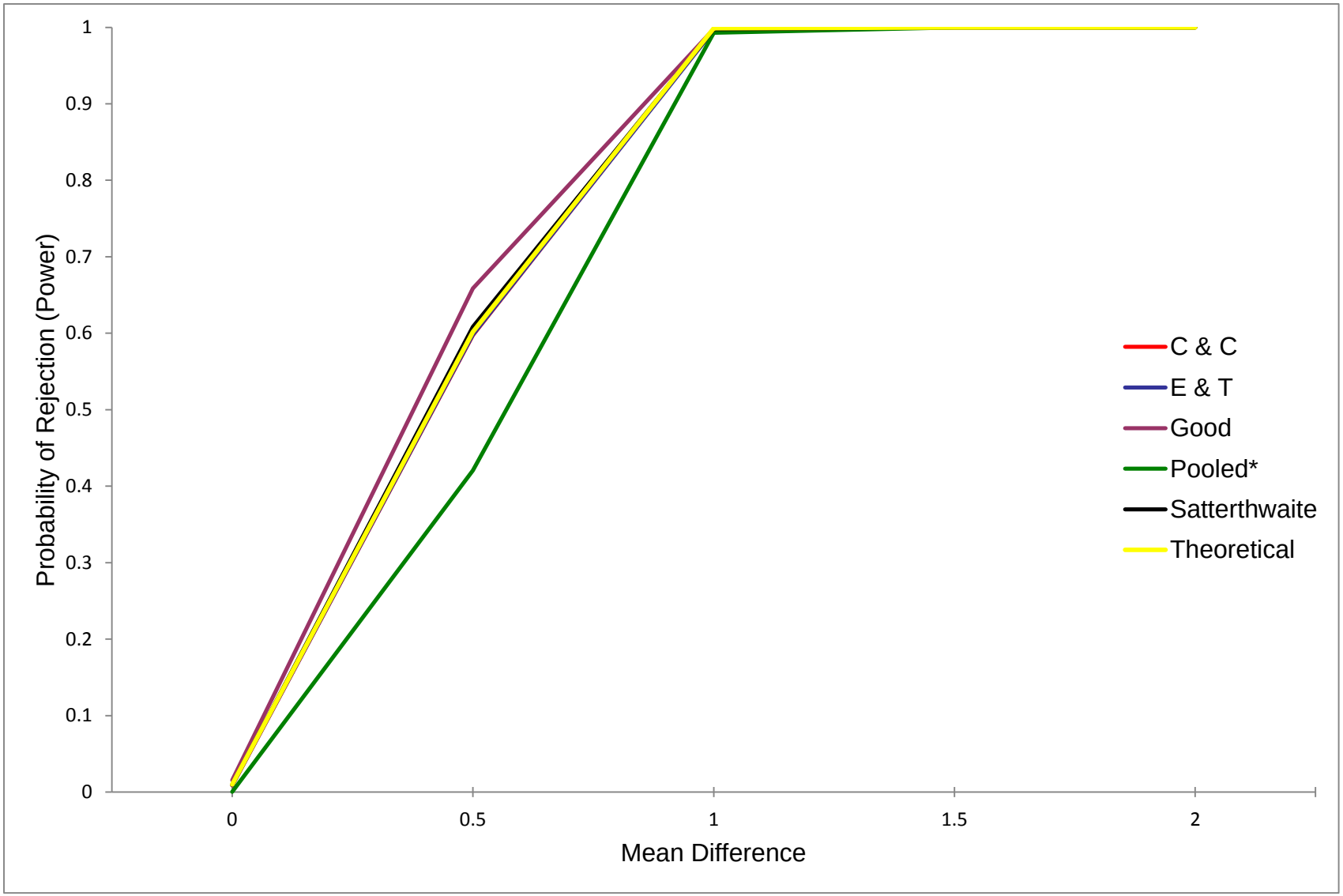

Figure B16. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

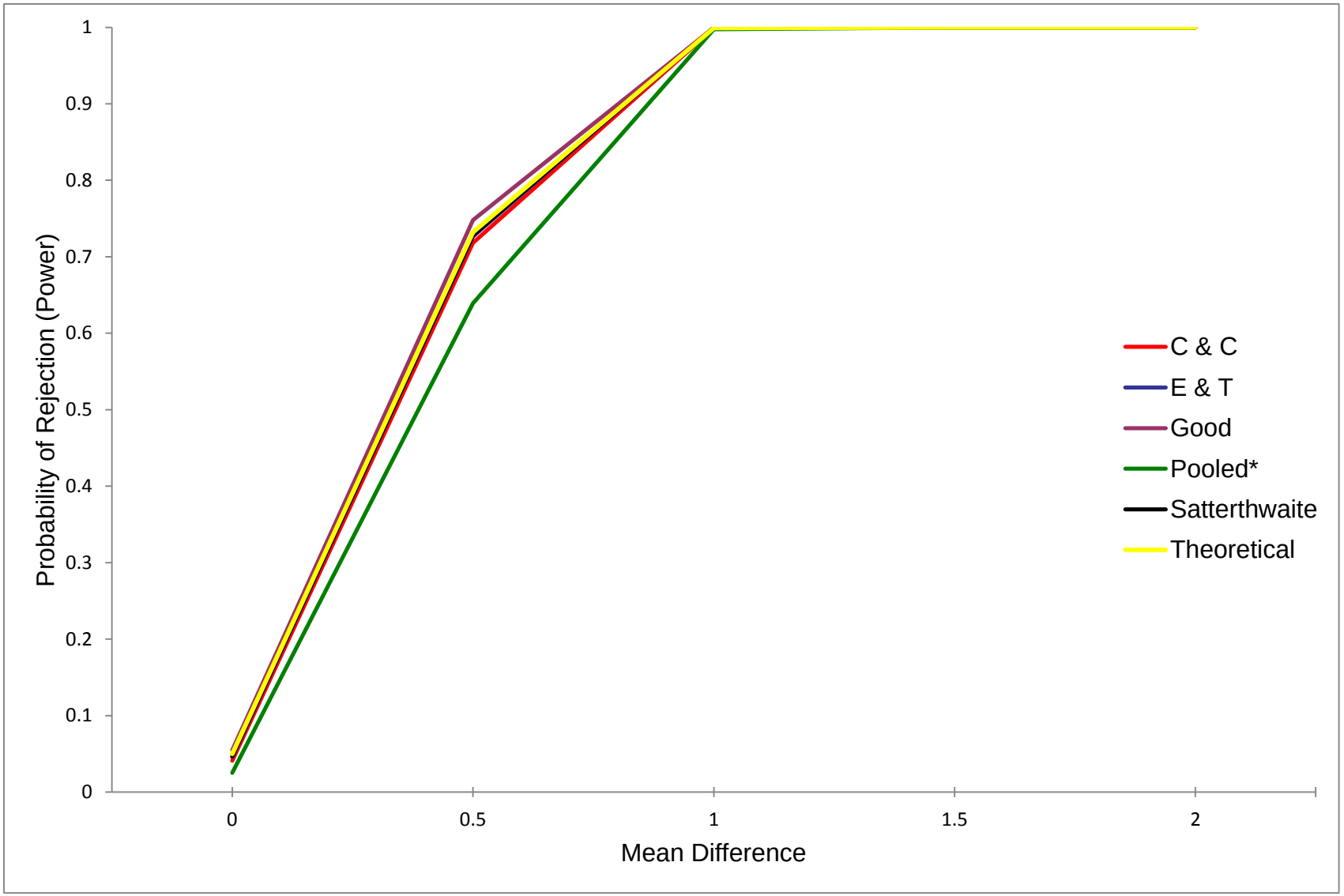

Figure B17. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

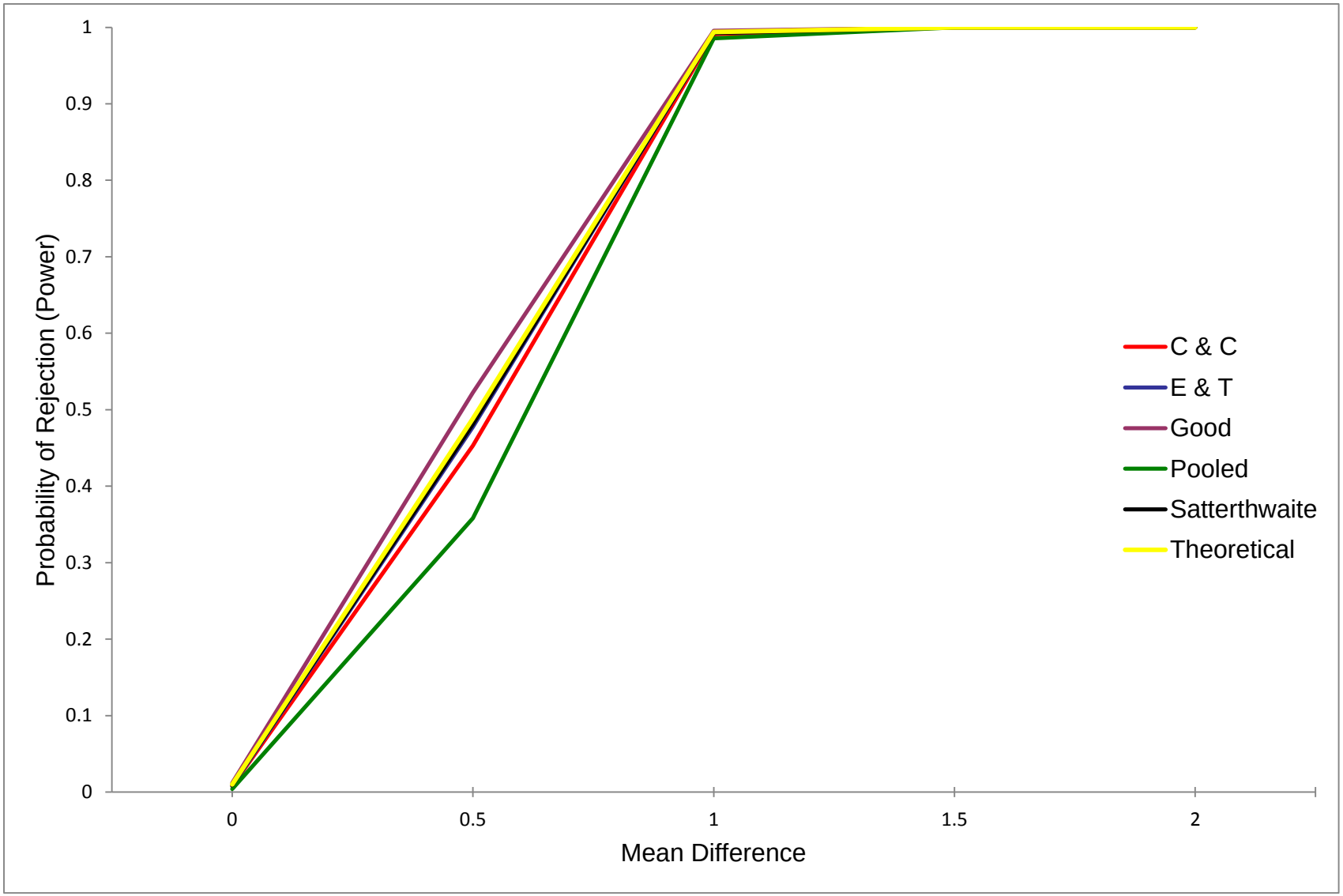

Figure B18. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

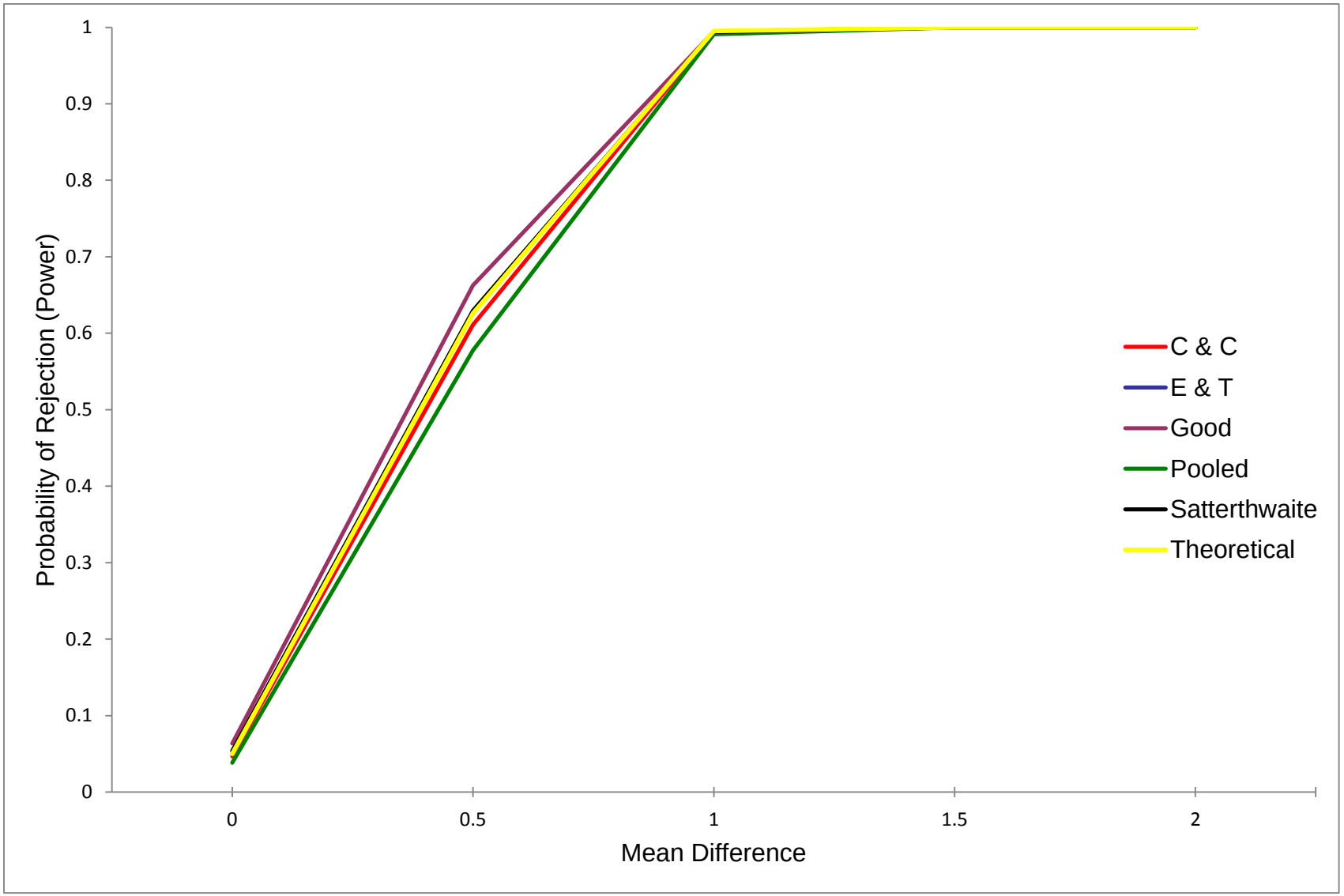

Figure B19. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

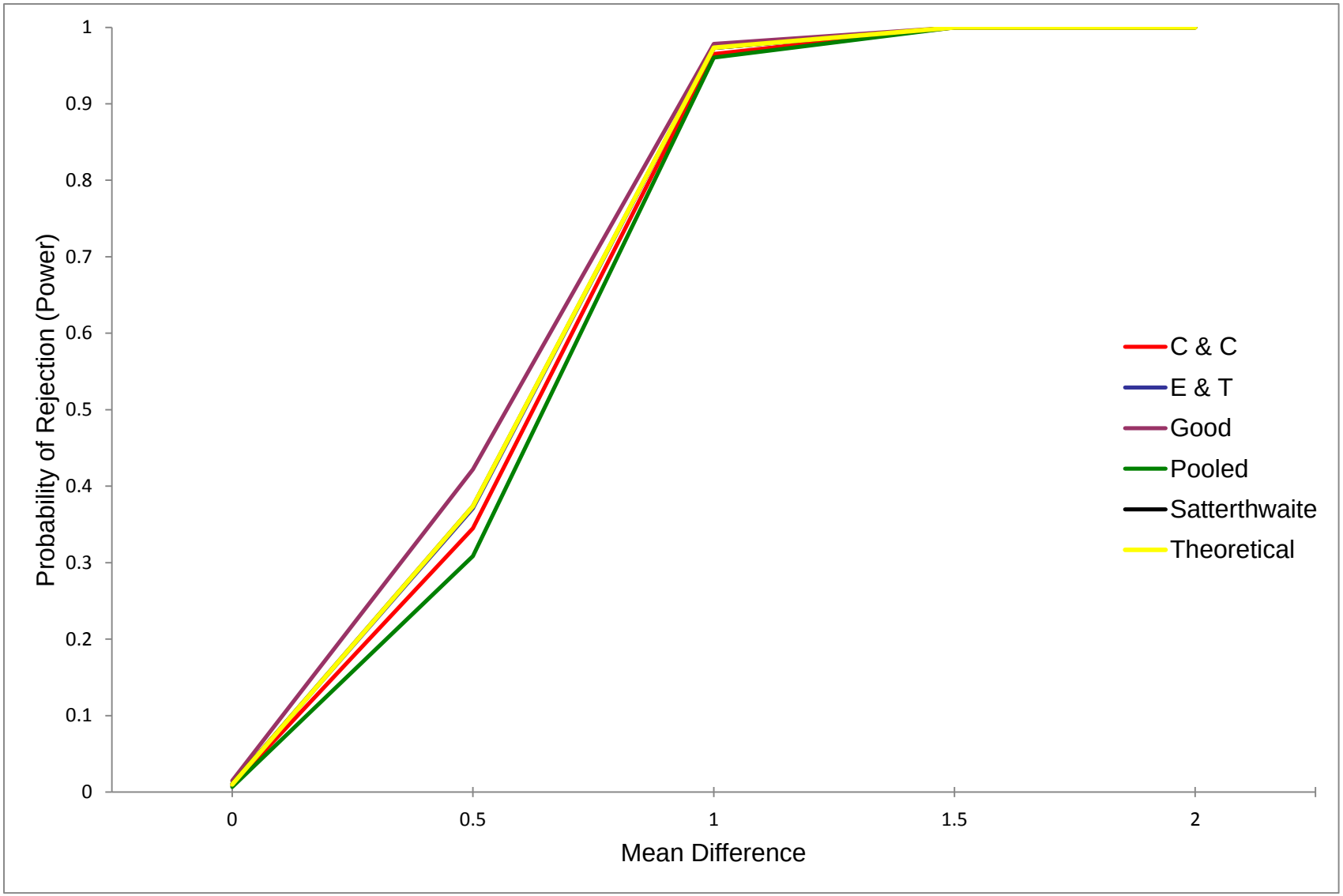

Figure B20. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

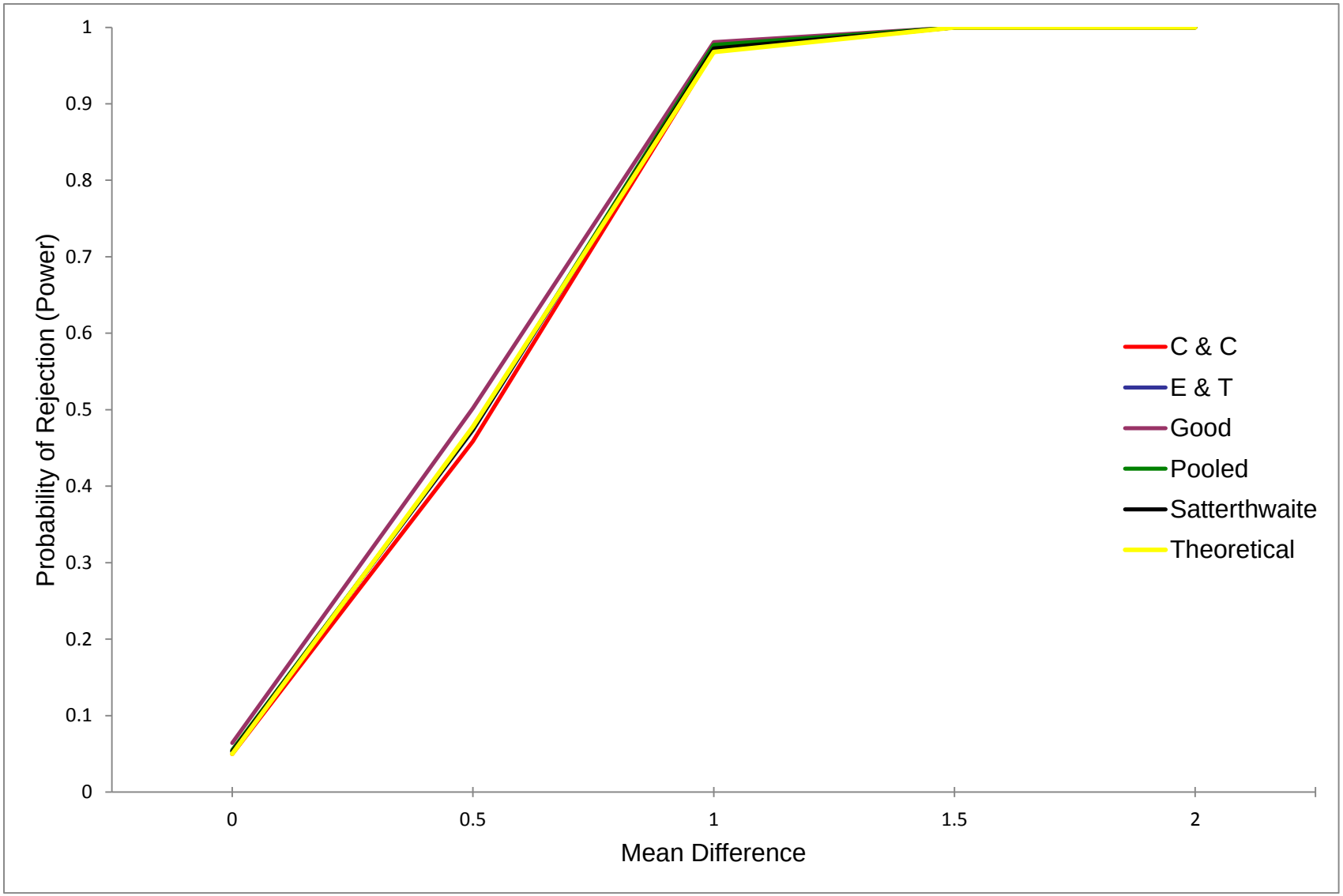

Figure B21. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

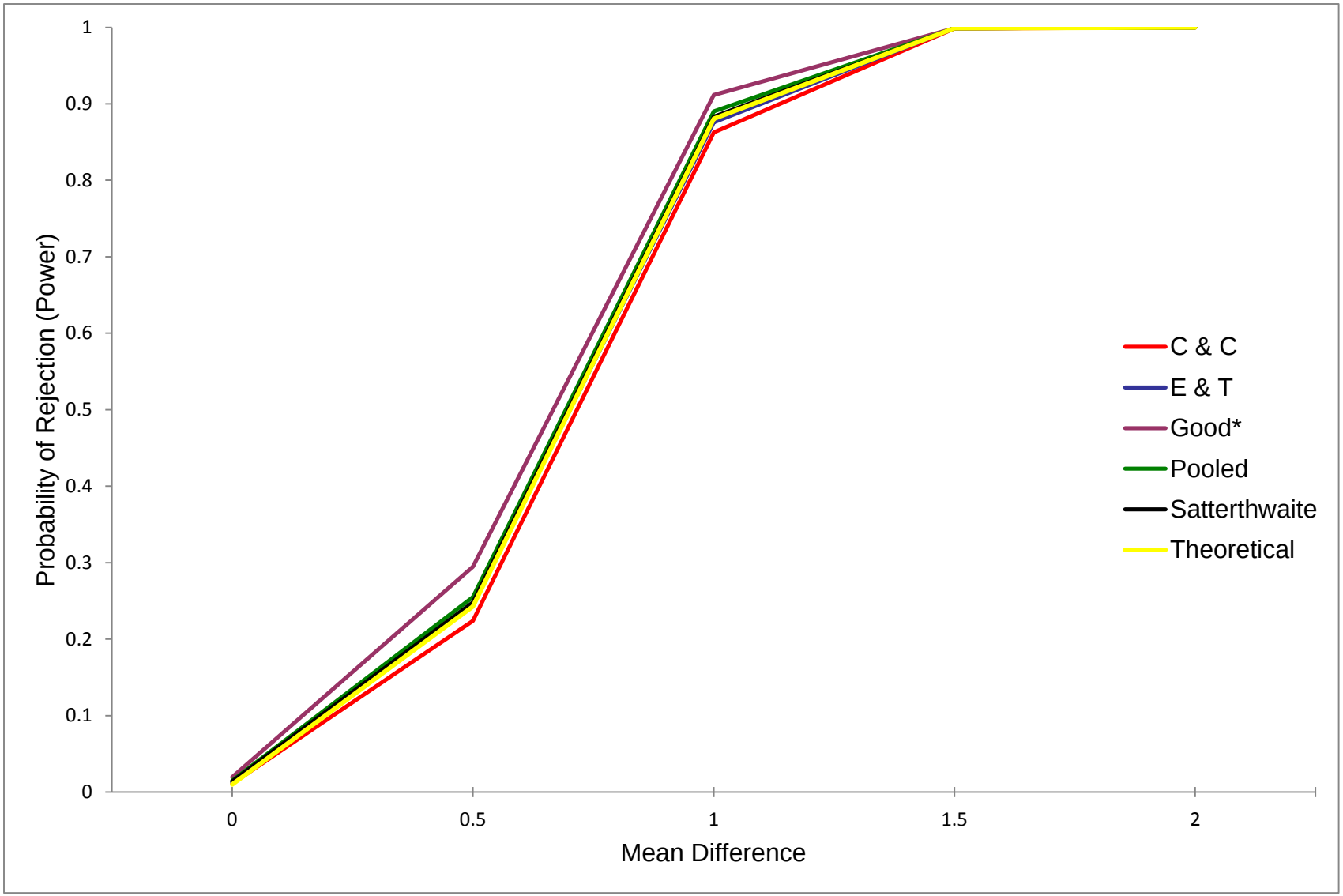

Figure B22. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

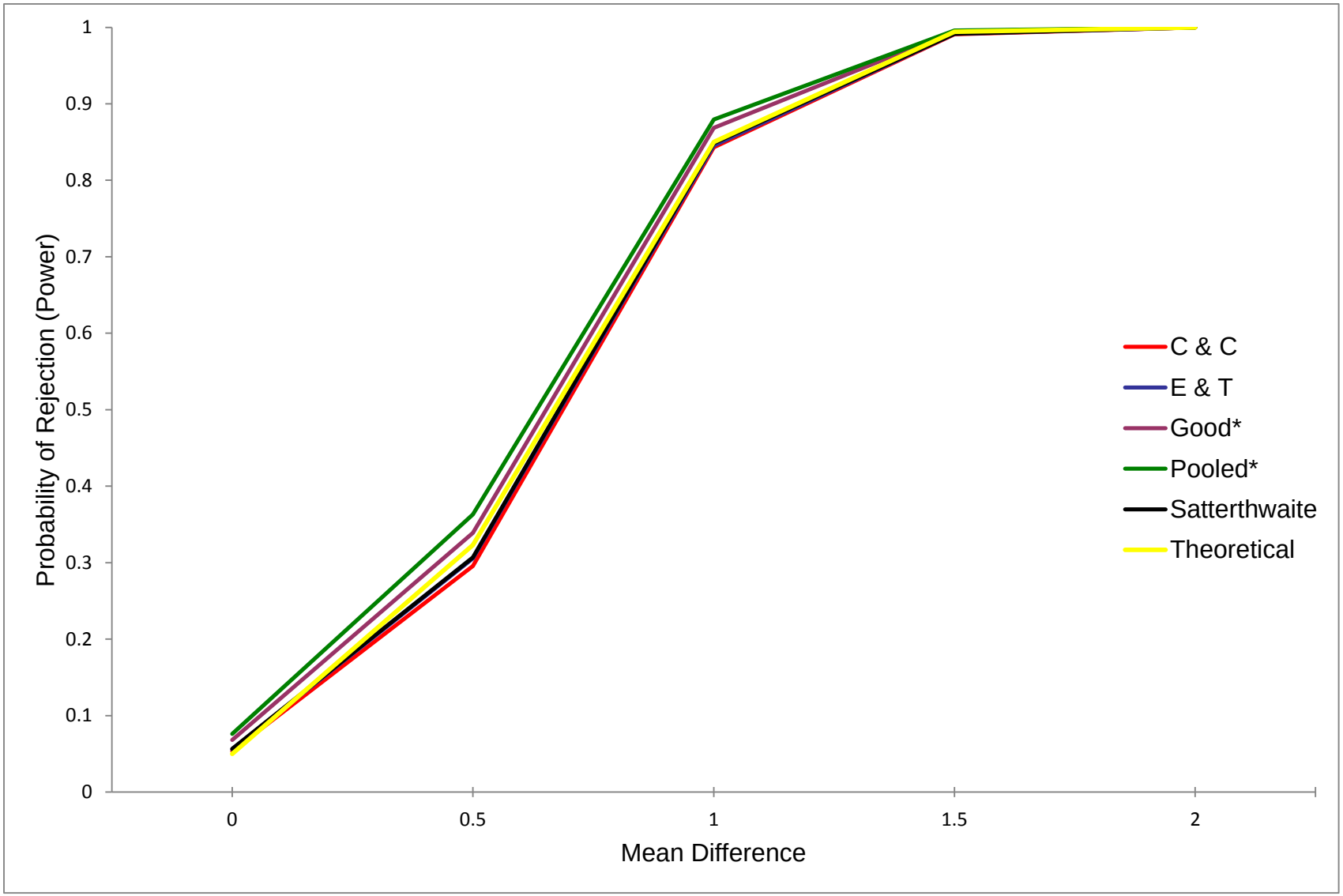

Figure B23. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

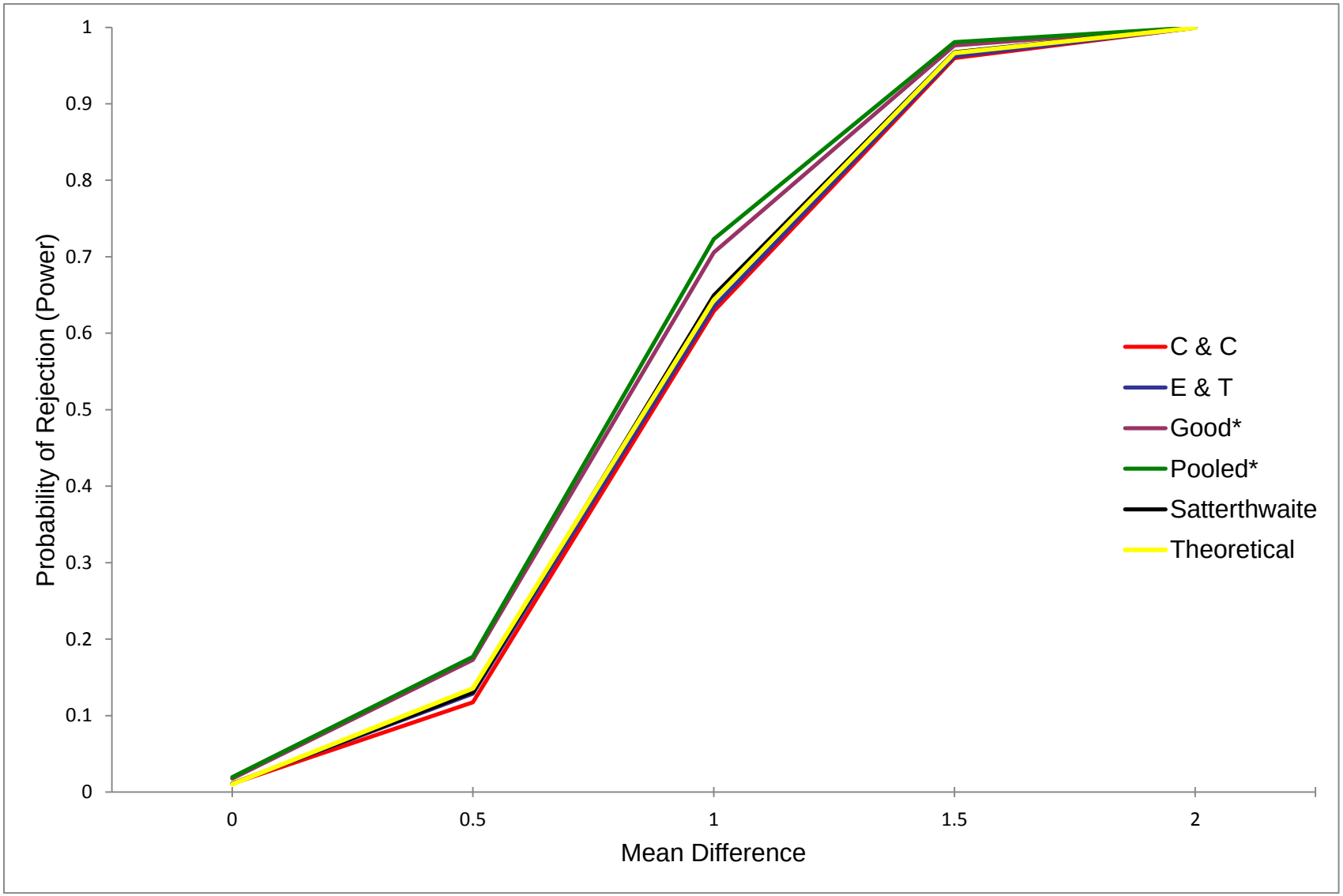

Figure B24. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

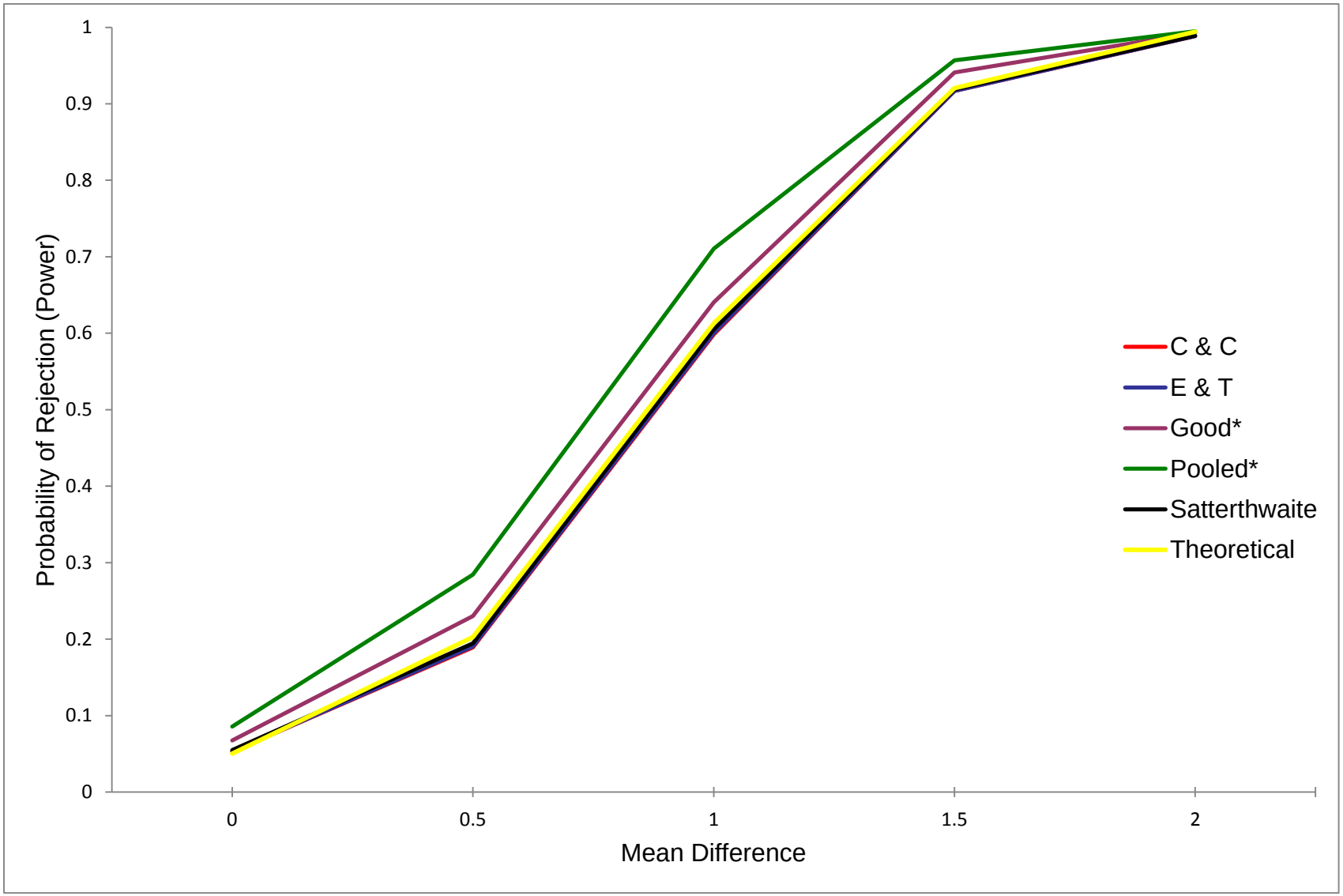

Figure B25. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

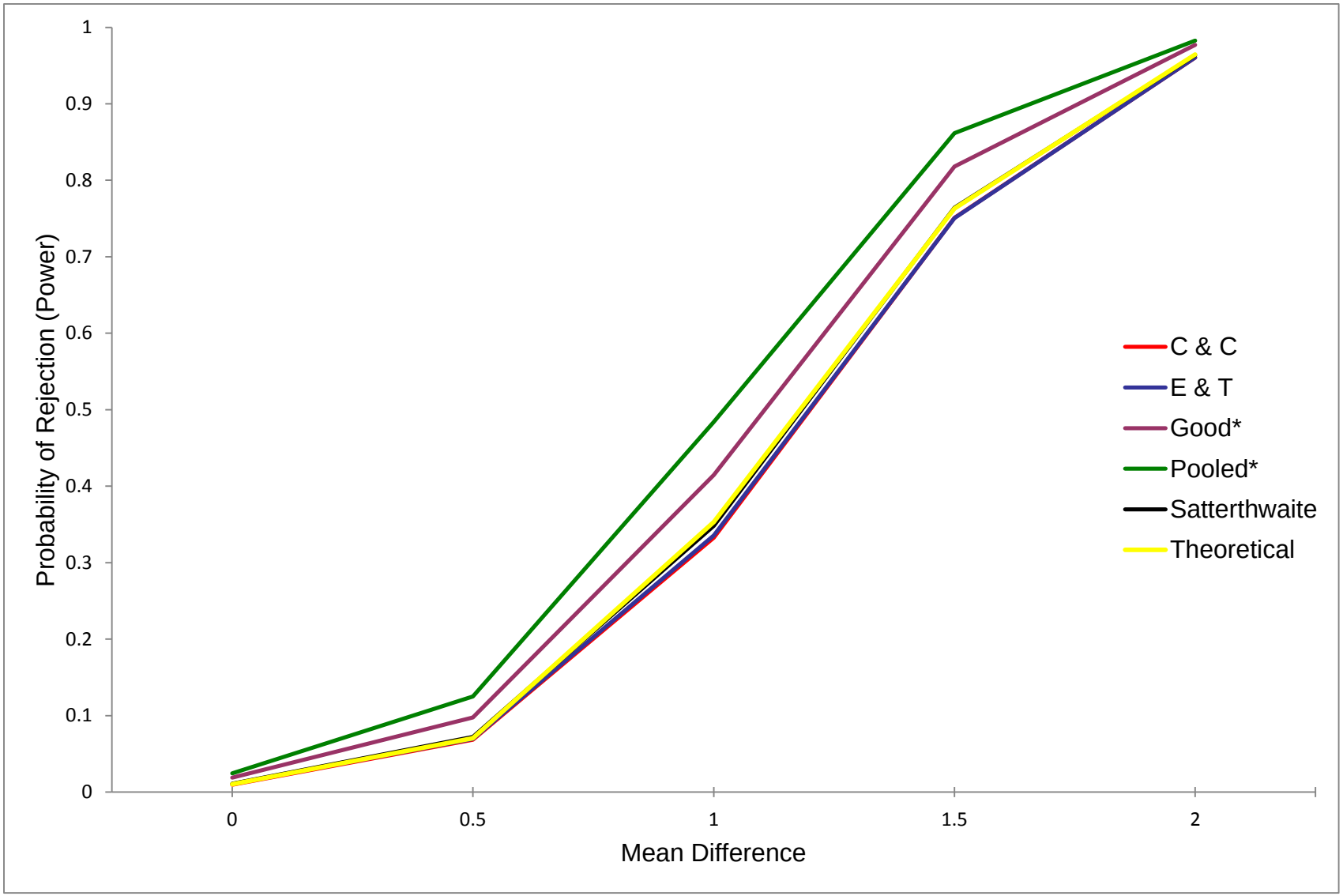

Figure B26. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

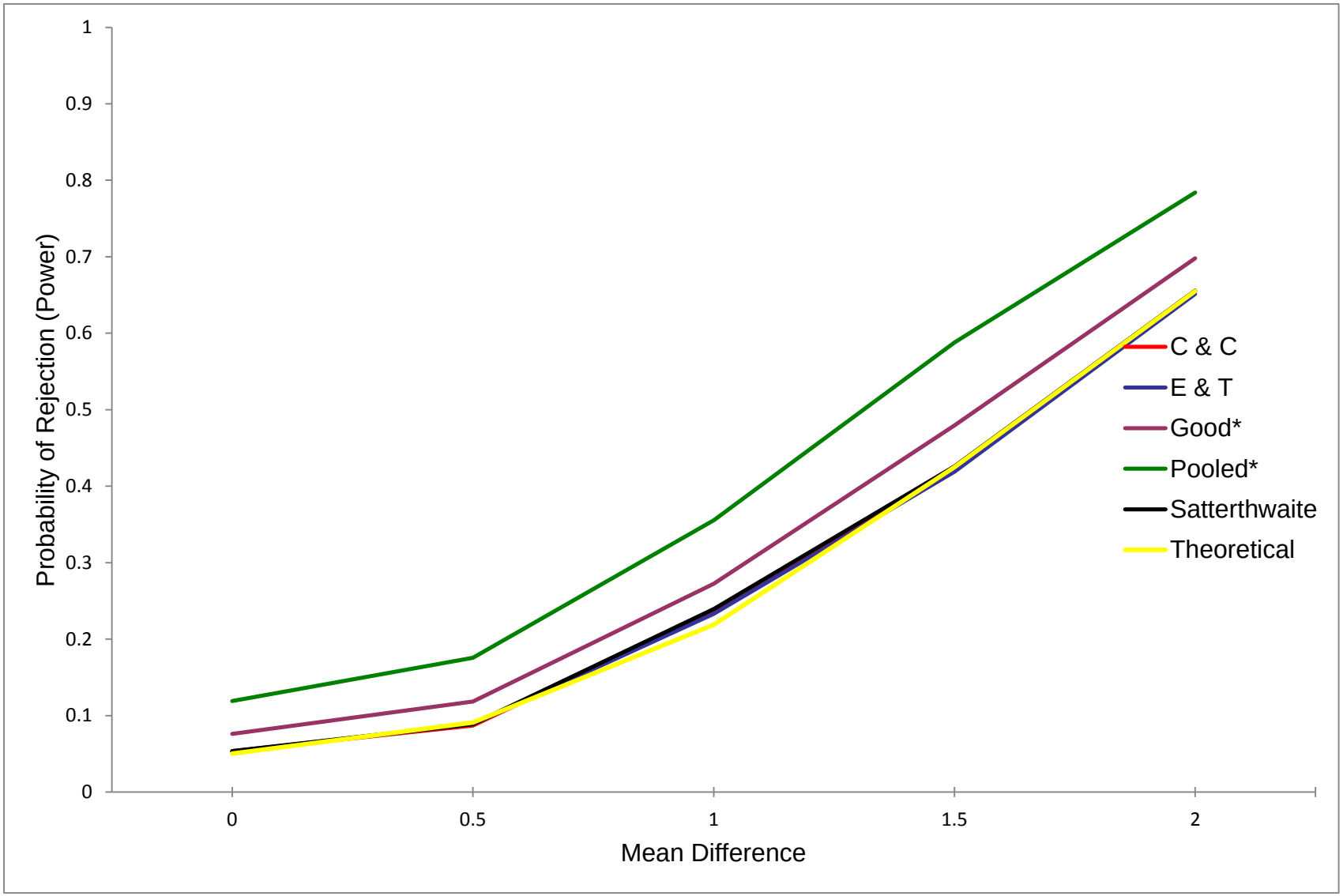

Figure B27. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

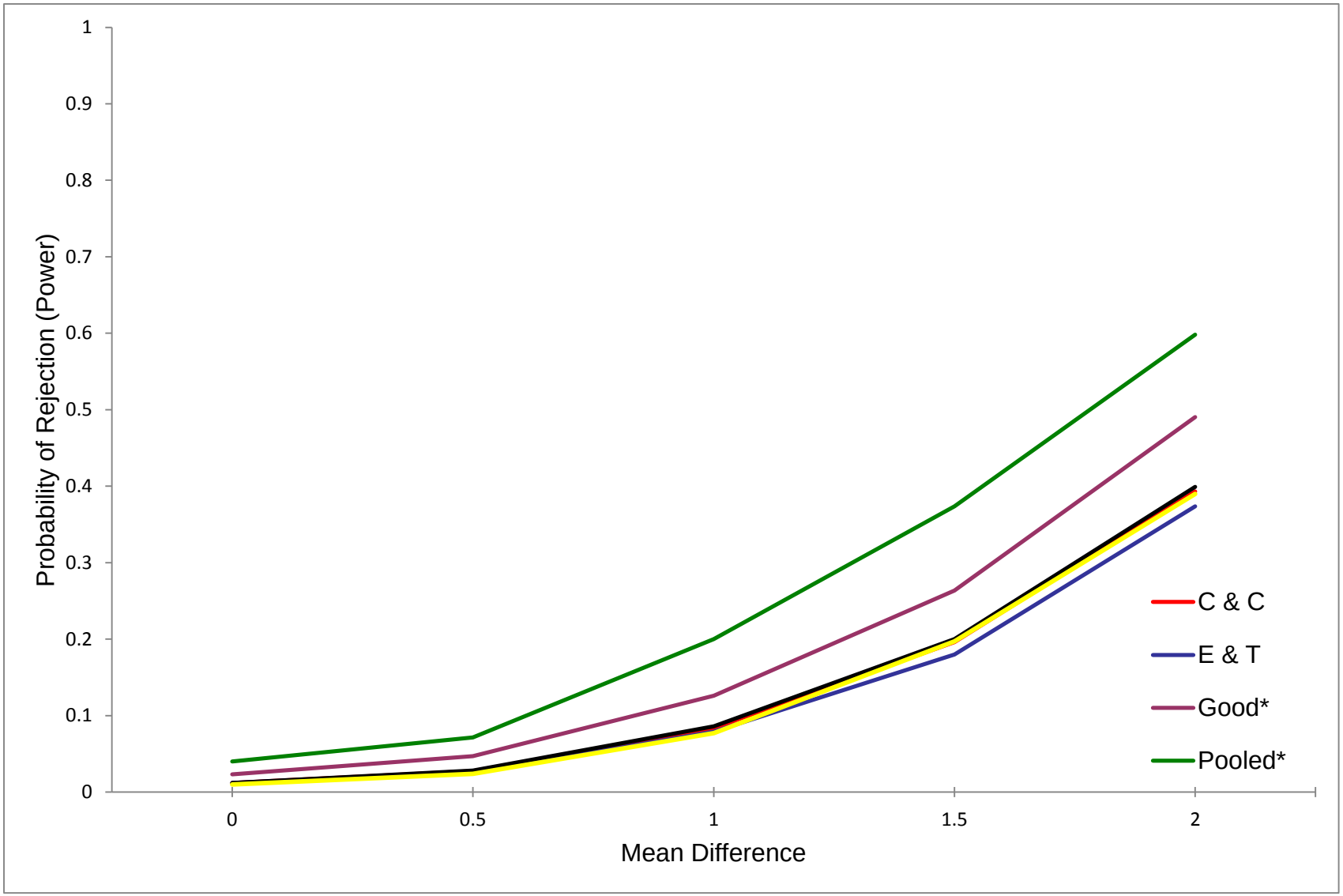

Figure B28. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 38$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 3.0 (i.e., $n_1 = 25$, $n_2 = 75$)

Table B19

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0400	0.0050
E & T	0.0430	0.0065
Good	0.0590	0.0150
Pooled	0.0015*	<.0005*
Satterthwaite	0.0445	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B20

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0405	0.0045
E & T	0.0525	0.0080
Good	0.0670*	0.0140
Pooled	0.0070*	0.0010*
Satterthwaite	0.0520	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B21

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0365	0.0045
E & T	0.0445	0.0055
Good	0.0640	0.0160
Pooled	0.0155*	0.0015*
Satterthwaite	0.0450	0.0065

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B22

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0375	0.0065
E & T	0.0445	0.0075
Good	0.0780*	0.0200*
Pooled	0.0445	0.0100
Satterthwaite	0.0465	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B23

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0420	0.0090
E & T	0.0425	0.0115
Good	0.0820*	0.0285*
Pooled	0.0990*	0.0275*
Satterthwaite	0.0465	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B24

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0465	0.0075
E & T	0.0440	0.0070
Good	0.0880*	0.0365*
Pooled	0.1535*	0.0620*
Satterthwaite	0.0490	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B25

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0415	0.0075
E & T	0.0295*	0.0050
Good	0.0845*	0.0350*
Pooled	0.2230*	0.1055*
Satterthwaite	0.0415	0.0080

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B26

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0485	0.0475	0.0500	0.0505	0.0490	0.0430	0.0515
	0.5	0.9690	0.8970	0.7600	0.5405	0.3350	0.2085	0.0900
	1.0	1.0000	1.0000	0.9995	0.9905	0.8895	0.6445	0.2115
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9345	0.4380
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9955	0.6375
Efron & Tibshirani	0.0	0.0490	0.0525	0.0530	0.0515	0.0485	0.0410	0.0500
	0.5	0.9690	0.9000	0.7715	0.5480	0.3355	0.2090	0.0875
	1.0	1.0000	1.0000	0.9995	0.9915	0.8885	0.6420	0.2065
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9340	0.4340
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9955	0.6340
Good	0.0	0.0545	0.0570	0.0620	0.0600	0.0605	0.0590	0.0670
	0.5	0.9720	0.9115	0.7920	0.5850	0.3810	0.2470	0.1200
	1.0	1.0000	1.0000	0.9995	0.9940	0.9065	0.6985	0.2530
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9510	0.4930
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9975	0.6805
Pooled	0.0	0.0025	0.0085	0.0235	0.0520	0.0925	0.1510	0.2210
	0.5	0.7750	0.7165	0.6440	0.5580	0.4795	0.4015	0.3080
	1.0	1.0000	1.0000	0.9985	0.9925	0.9400	0.8355	0.5020
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9810	0.7390
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8730
Satterthwaite	0.0	0.0485	0.0505	0.0540	0.0525	0.0505	0.0430	0.0520
	0.5	0.9705	0.9000	0.7710	0.5495	0.3390	0.2115	0.0900
	1.0	1.0000	1.0000	0.9995	0.9925	0.8900	0.6480	0.2125
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9350	0.4390
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9955	0.6390

Table B27

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 75$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0095	0.0070	0.0100	0.0085	0.0070	0.0040	0.0105
	0.5	0.9065	0.7170	0.4905	0.2885	0.1390	0.0605	0.0245
	1.0	1.0000	1.0000	0.9960	0.9385	0.6885	0.3745	0.0800
	1.5	1.0000	1.0000	1.0000	1.0000	0.9835	0.7835	0.2090
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9680	0.3780
Efron & Tibshirani	0.0	0.0110	0.0095	0.0100	0.0100	0.0075	0.0035	0.0085
	0.5	0.9110	0.7375	0.5075	0.3010	0.1370	0.0570	0.0225
	1.0	1.0000	1.0000	0.9965	0.9395	0.6875	0.3645	0.0740
	1.5	1.0000	1.0000	1.0000	1.0000	0.9825	0.7685	0.1905
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9630	0.3560
Good	0.0	0.0120	0.0130	0.0130	0.0180	0.0165	0.0095	0.0225
	0.5	0.9205	0.7640	0.5685	0.3655	0.1955	0.1050	0.0455
	1.0	1.0000	1.0000	0.9975	0.9585	0.7755	0.4715	0.1160
	1.5	1.0000	1.0000	1.0000	1.0000	0.9935	0.8475	0.2725
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9805	0.4760
Pooled	0.0	<.0005	<.0005	0.0040	0.0105	0.0300	0.0500	0.1080
	0.5	0.4035	0.3975	0.3585	0.3235	0.2790	0.2330	0.1835
	1.0	1.0000	0.9990	0.9870	0.9565	0.8535	0.6910	0.3545
	1.5	1.0000	1.0000	1.0000	1.0000	0.9965	0.9535	0.5990
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9975	0.7815
Satterthwaite	0.0	0.0100	0.0100	0.0100	0.0120	0.0080	0.0045	0.0105
	0.5	0.9125	0.7380	0.5150	0.3100	0.1460	0.0655	0.0250
	1.0	1.0000	1.0000	0.9965	0.9450	0.7040	0.3830	0.0800
	1.5	1.0000	1.0000	1.0000	1.0000	0.9855	0.7920	0.2105
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9690	0.3795

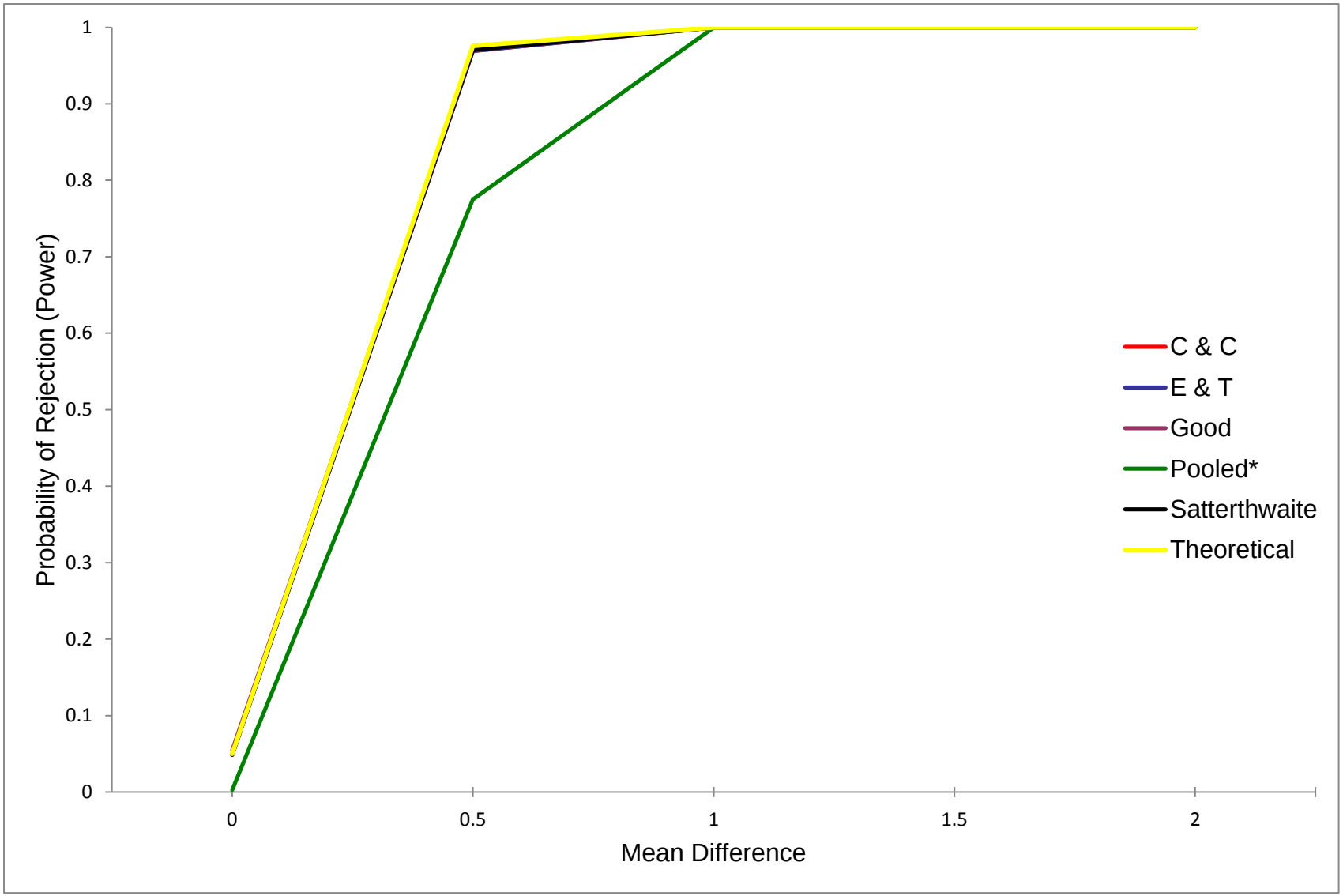

Figure B29. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

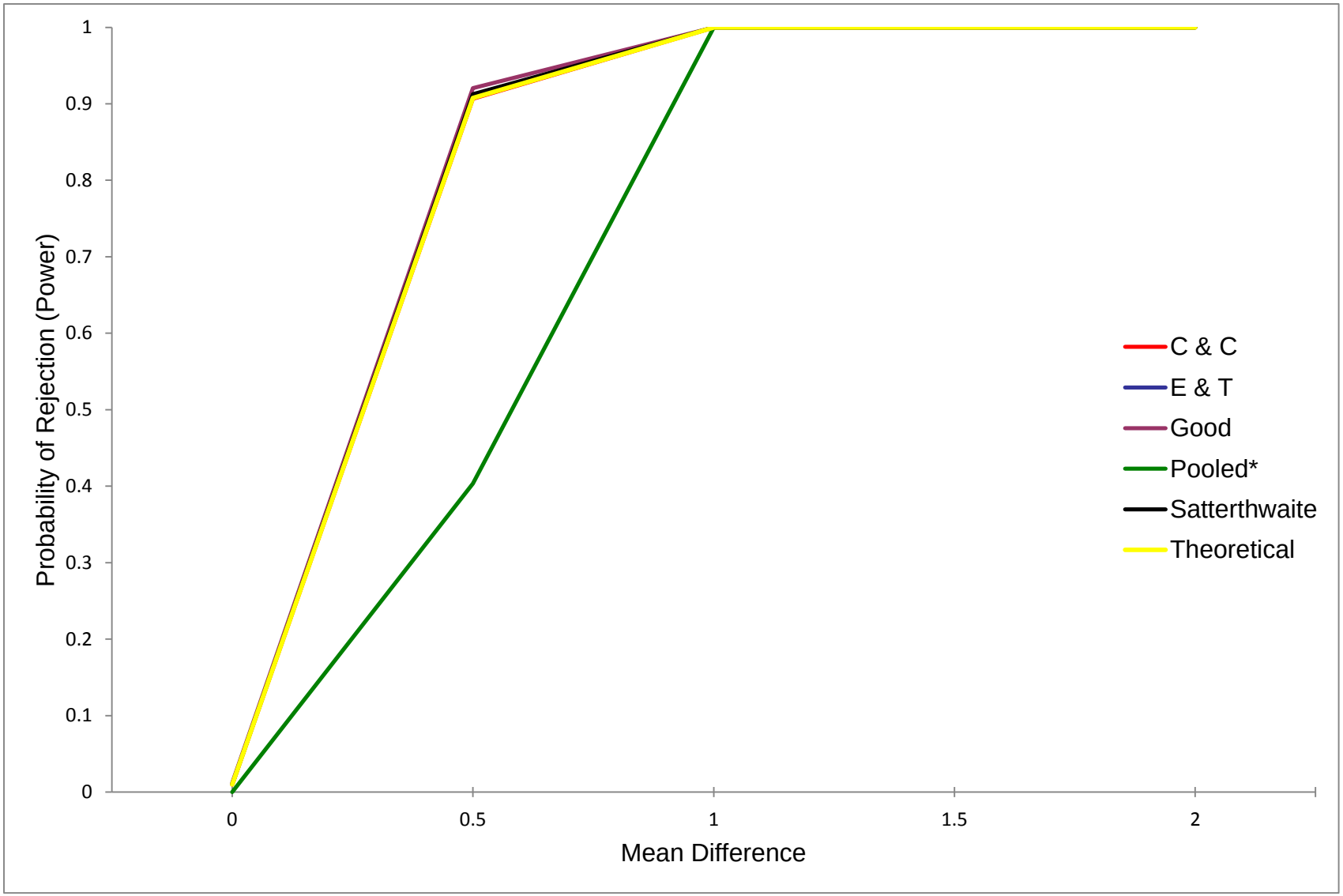

Figure B30. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

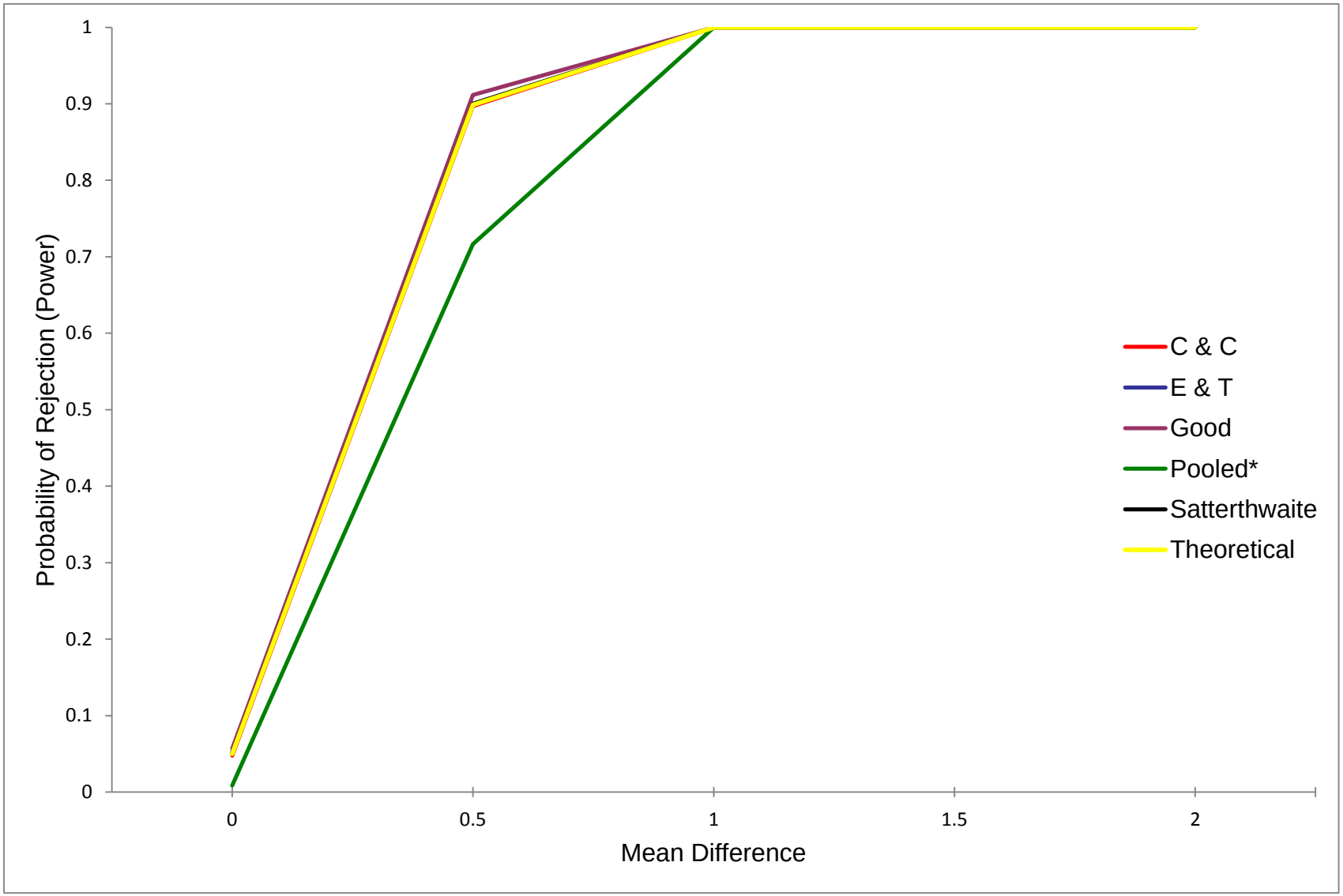

Figure B31. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

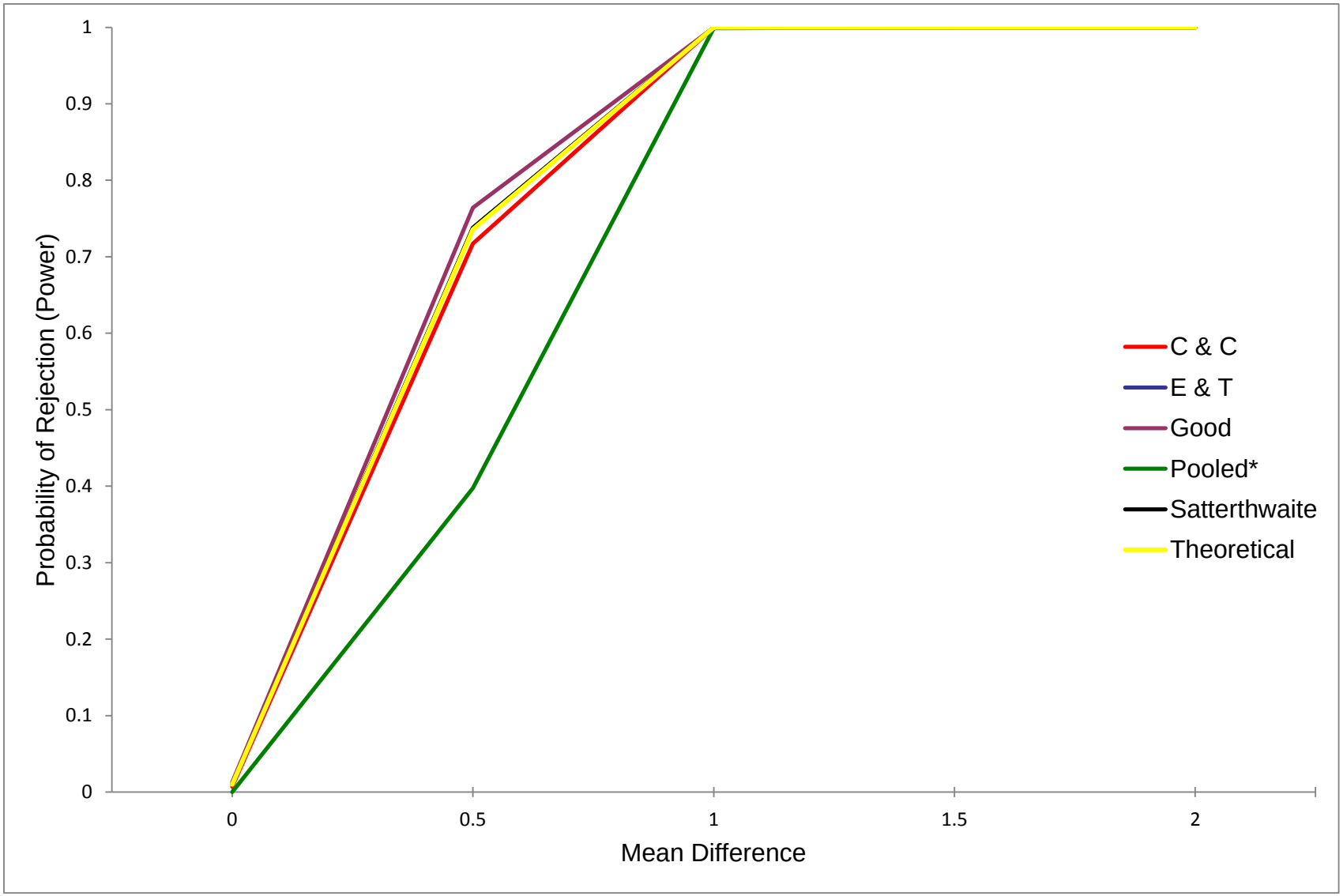

Figure B32. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

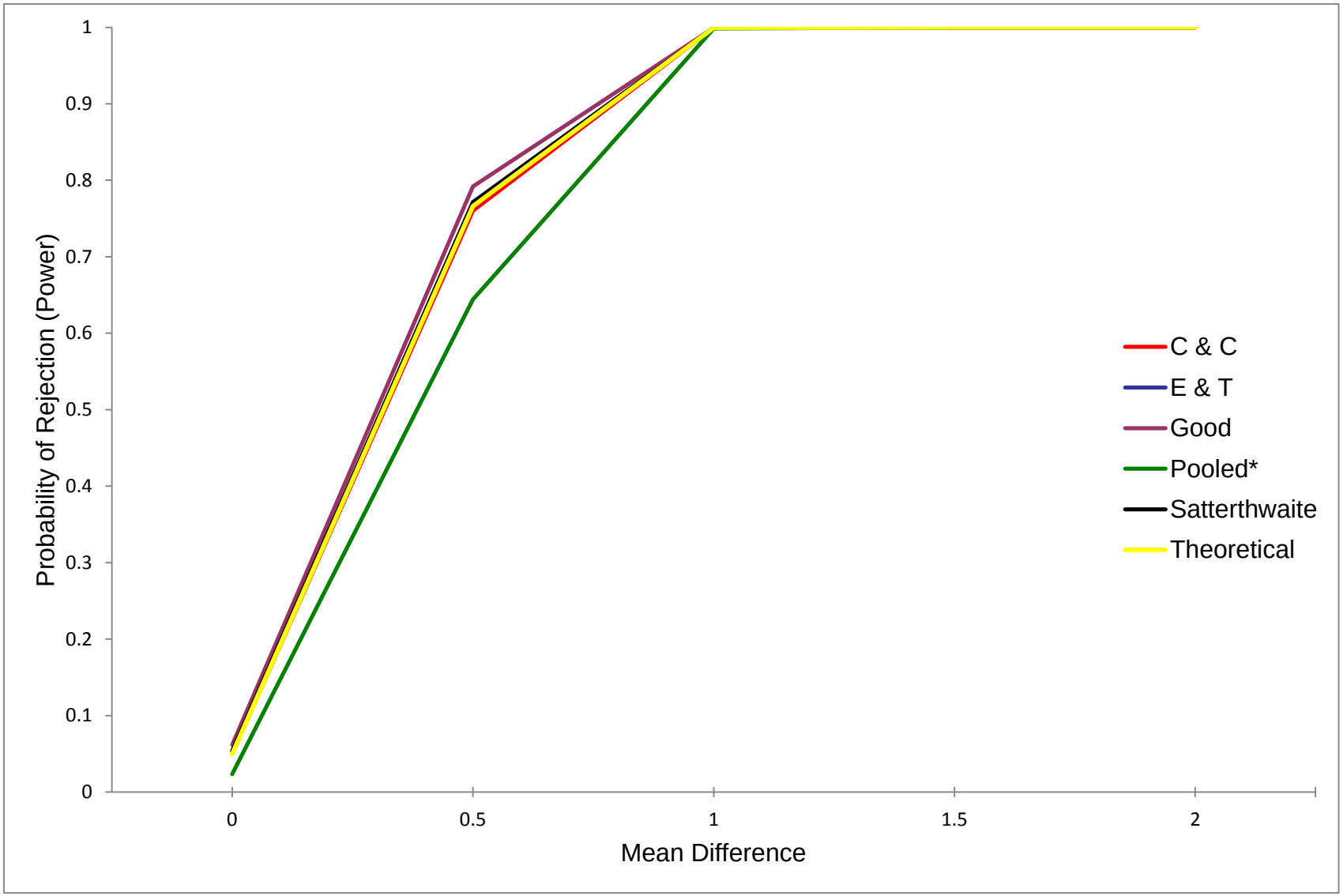

Figure B33. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

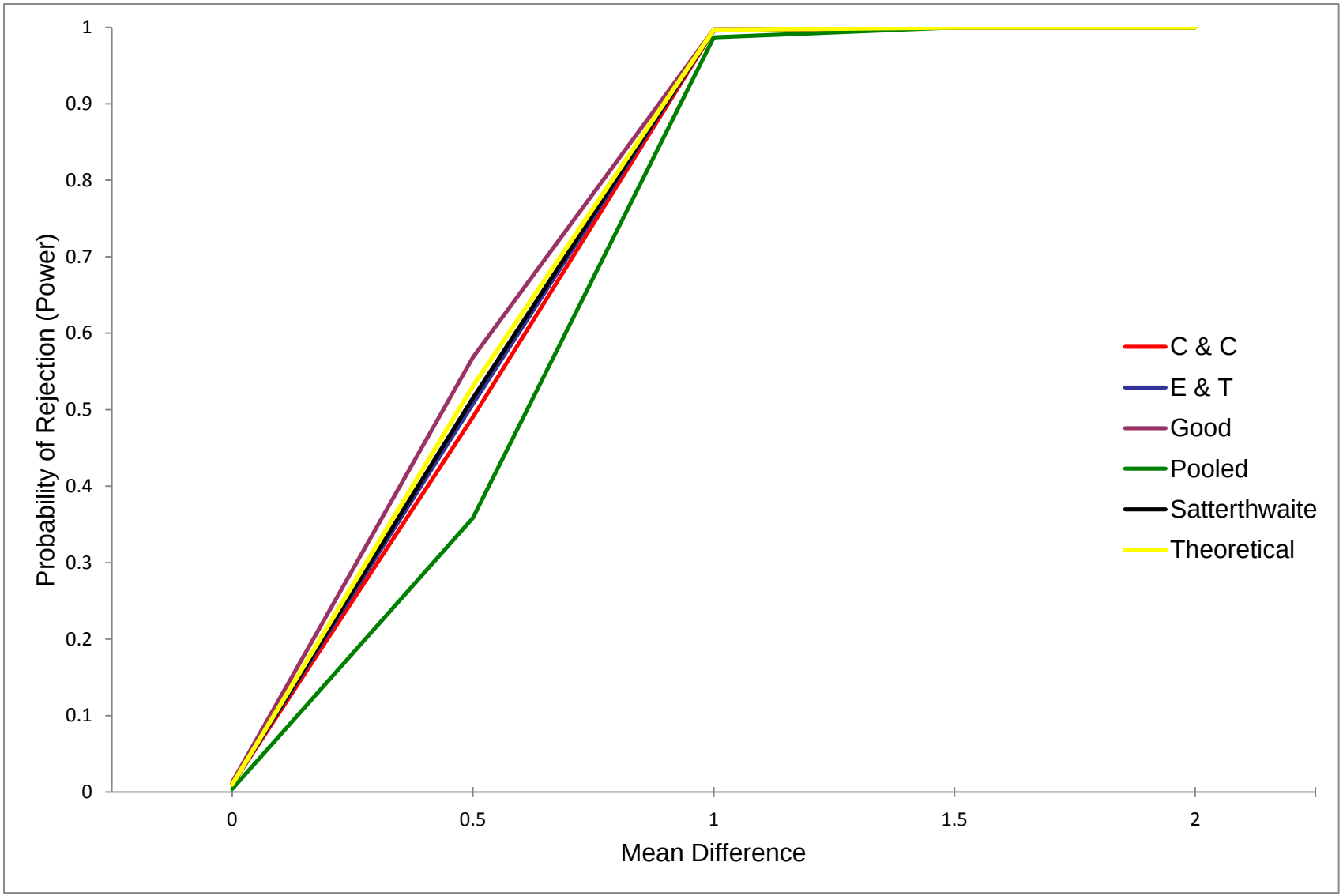

Figure B34. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

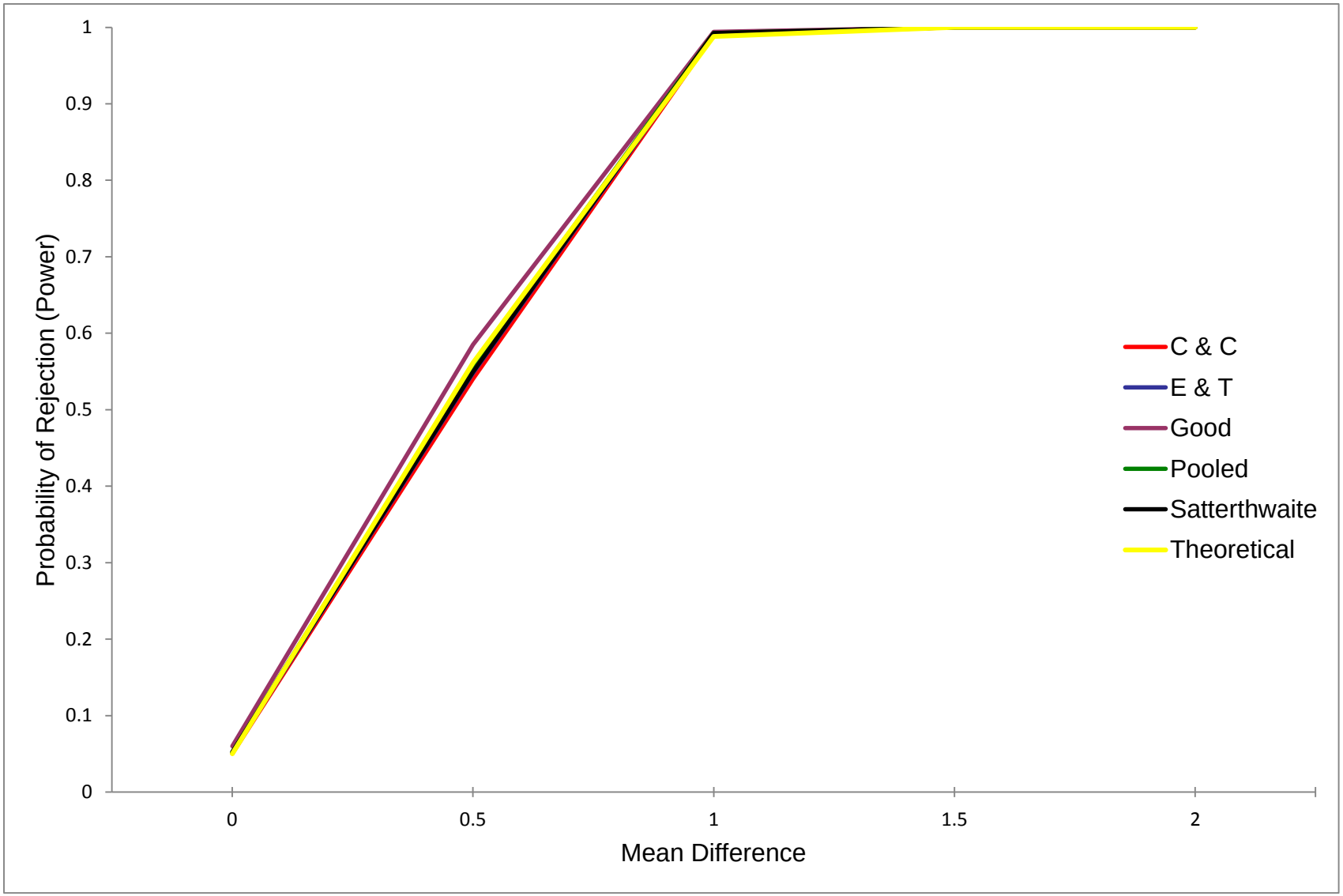

Figure B35. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

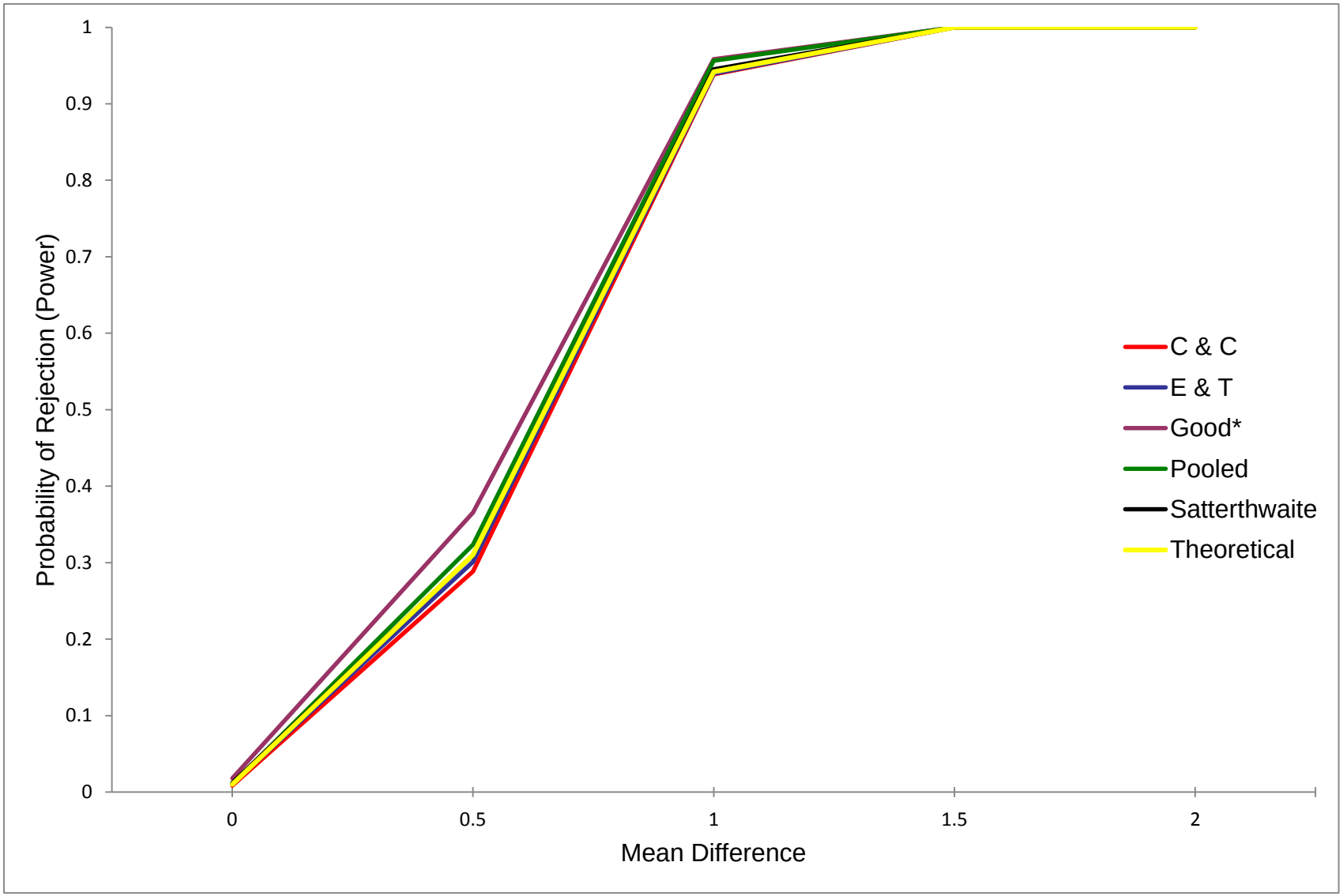

Figure B36. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

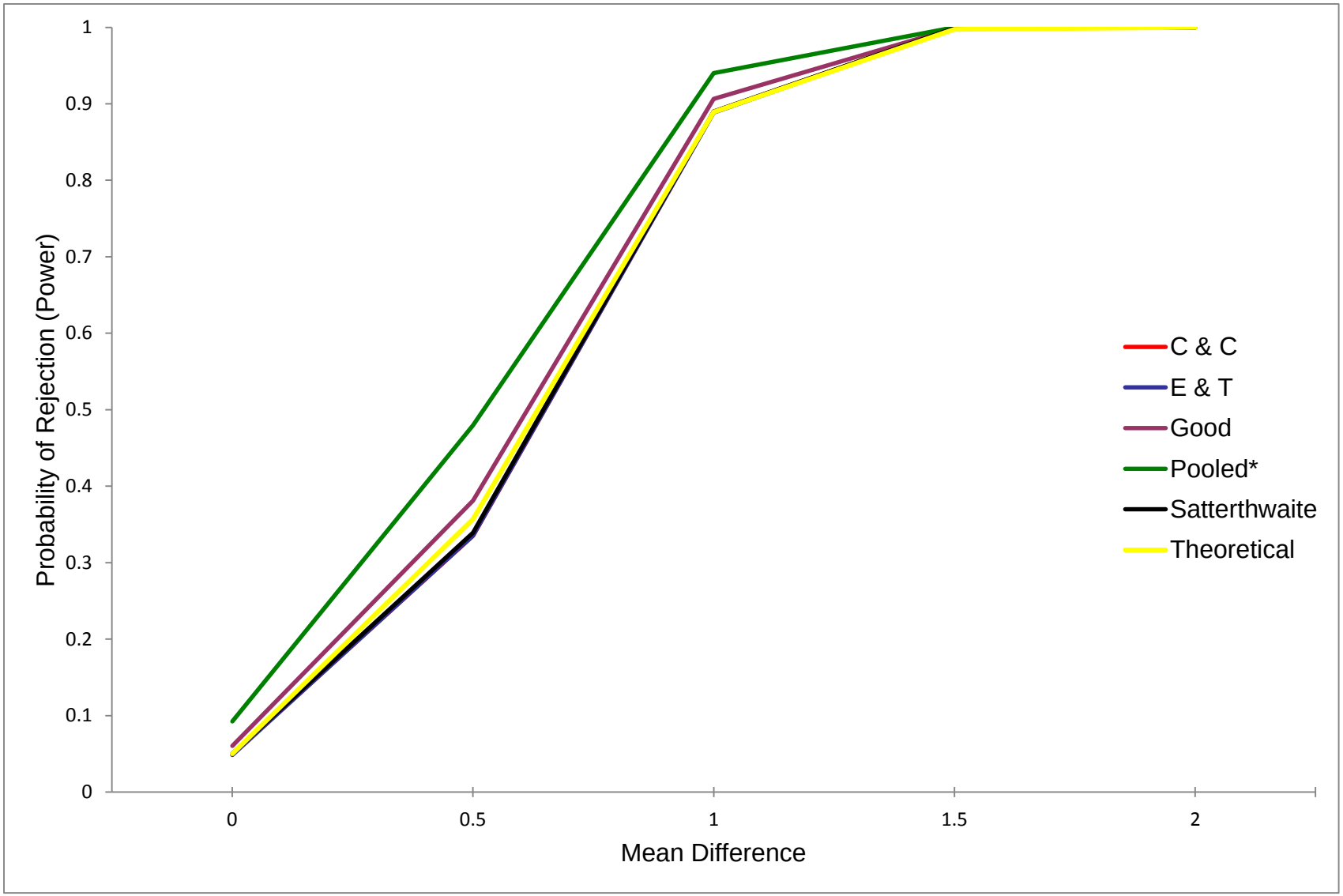

Figure B37. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

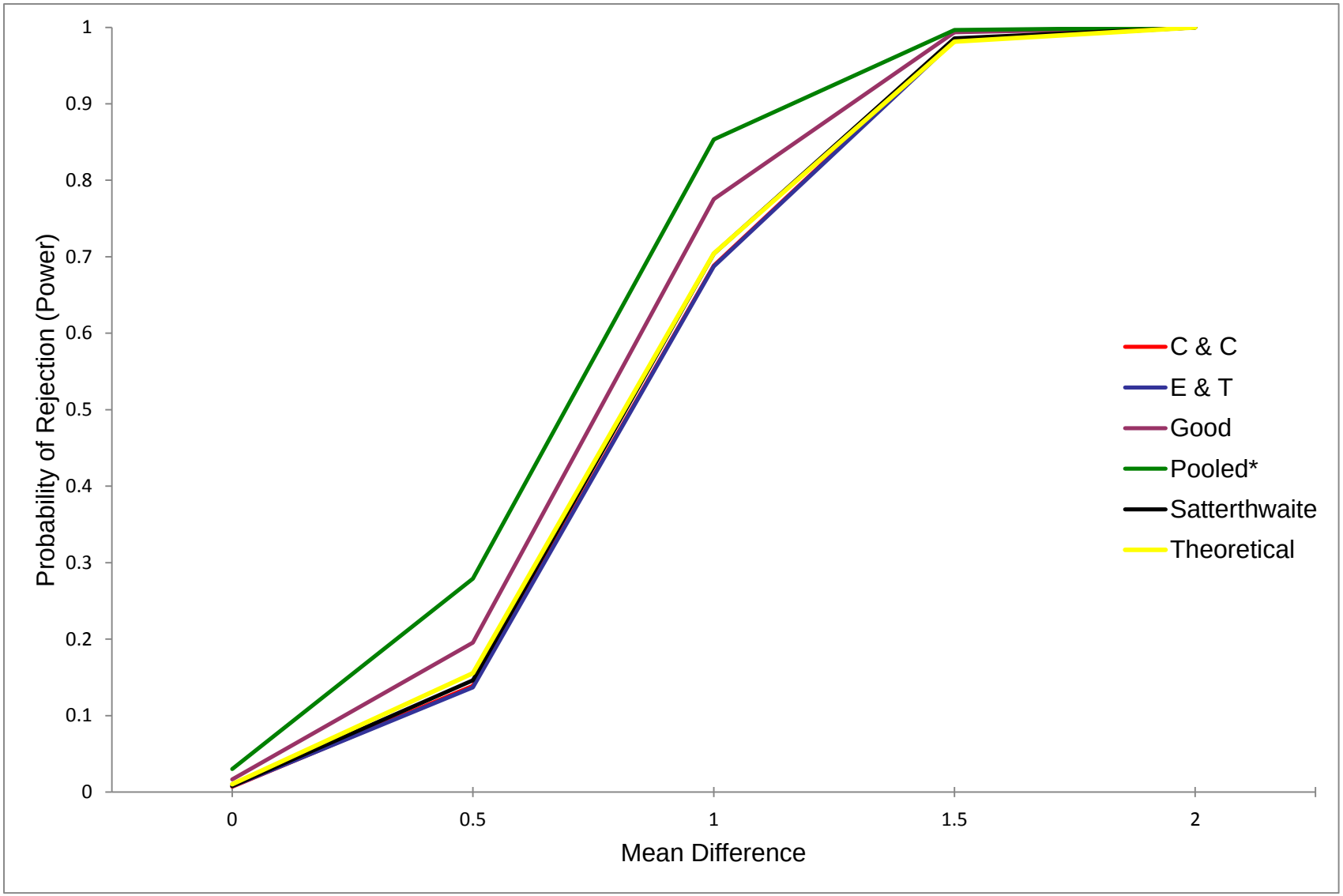

Figure B38. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

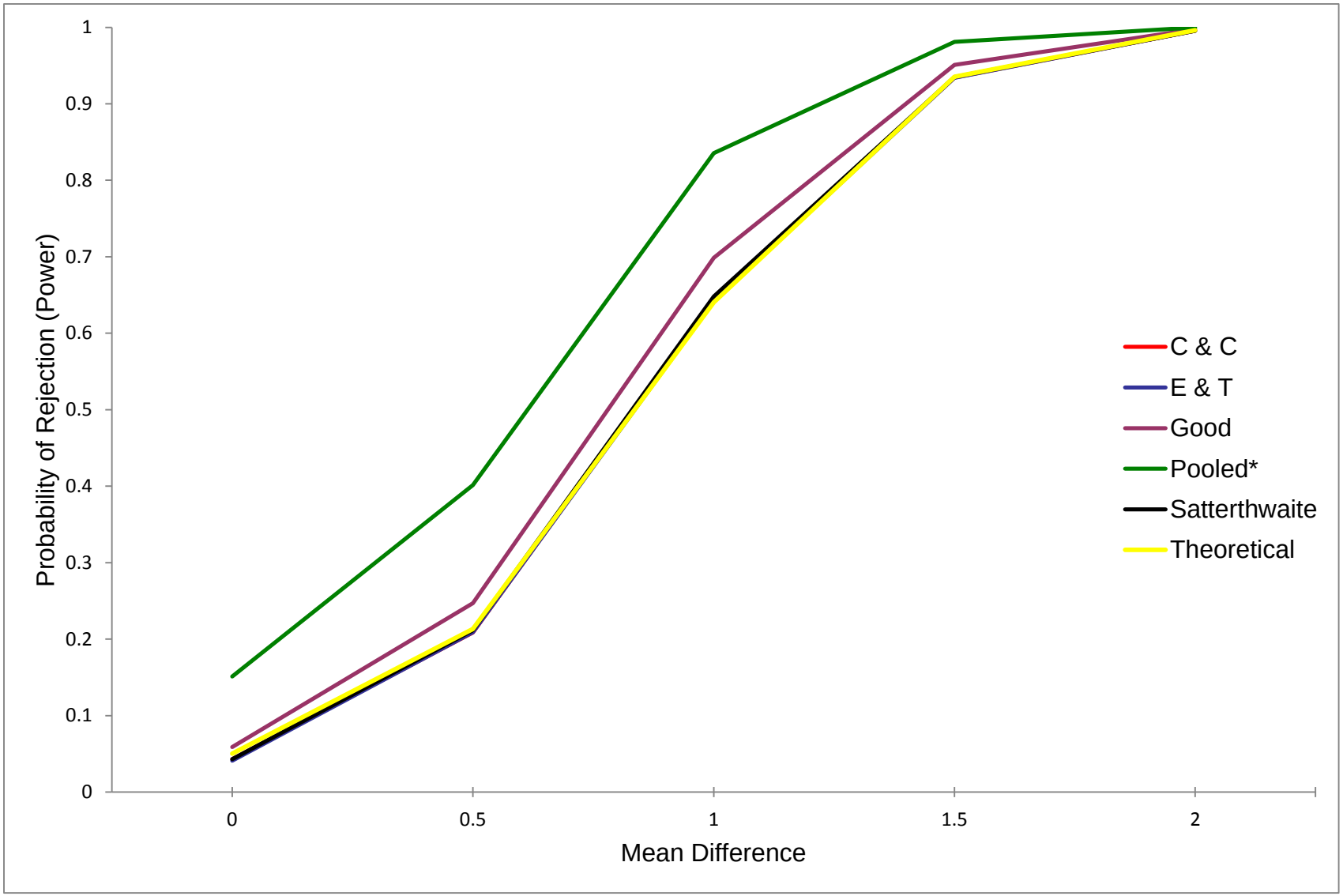

Figure B39. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

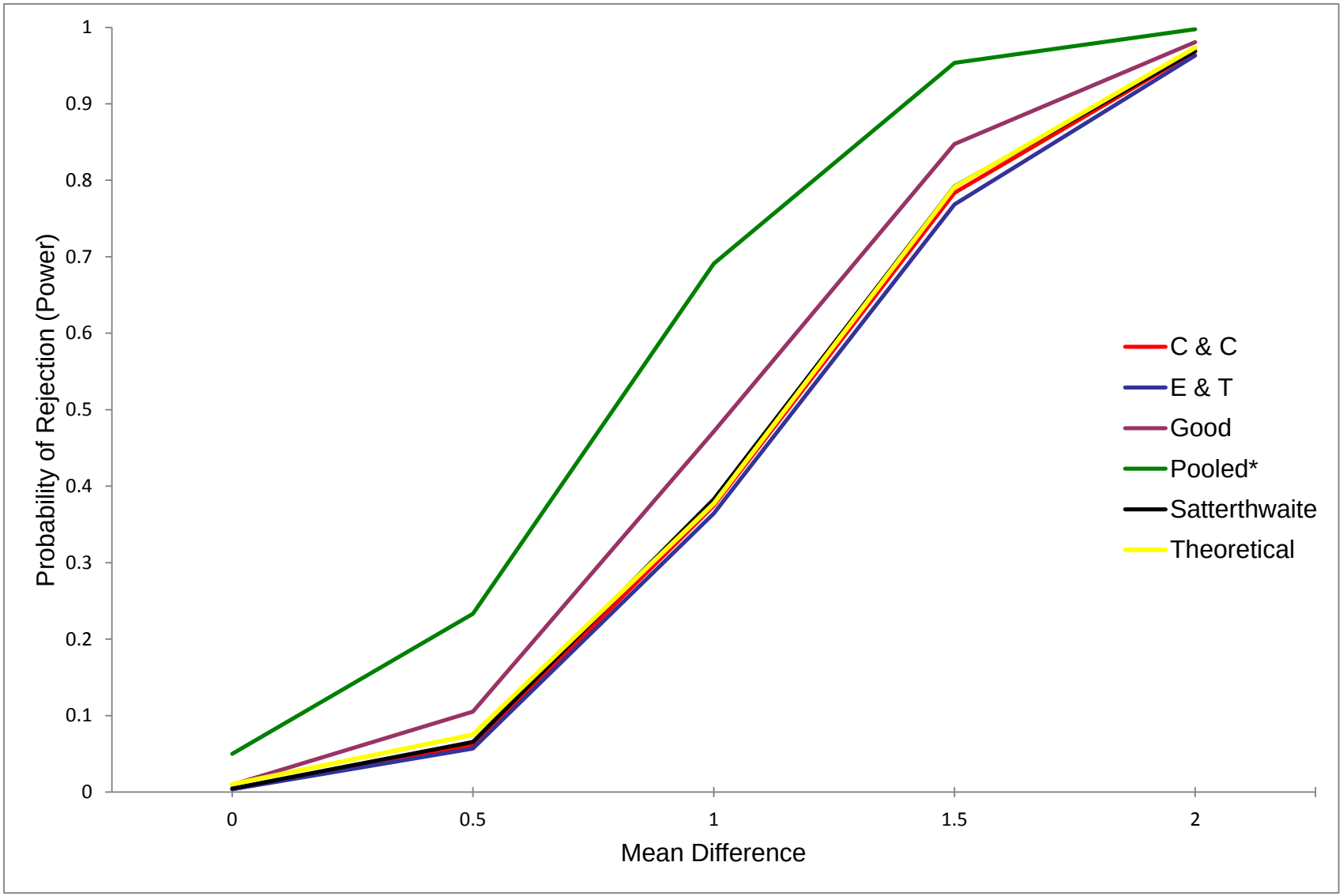

Figure B40. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

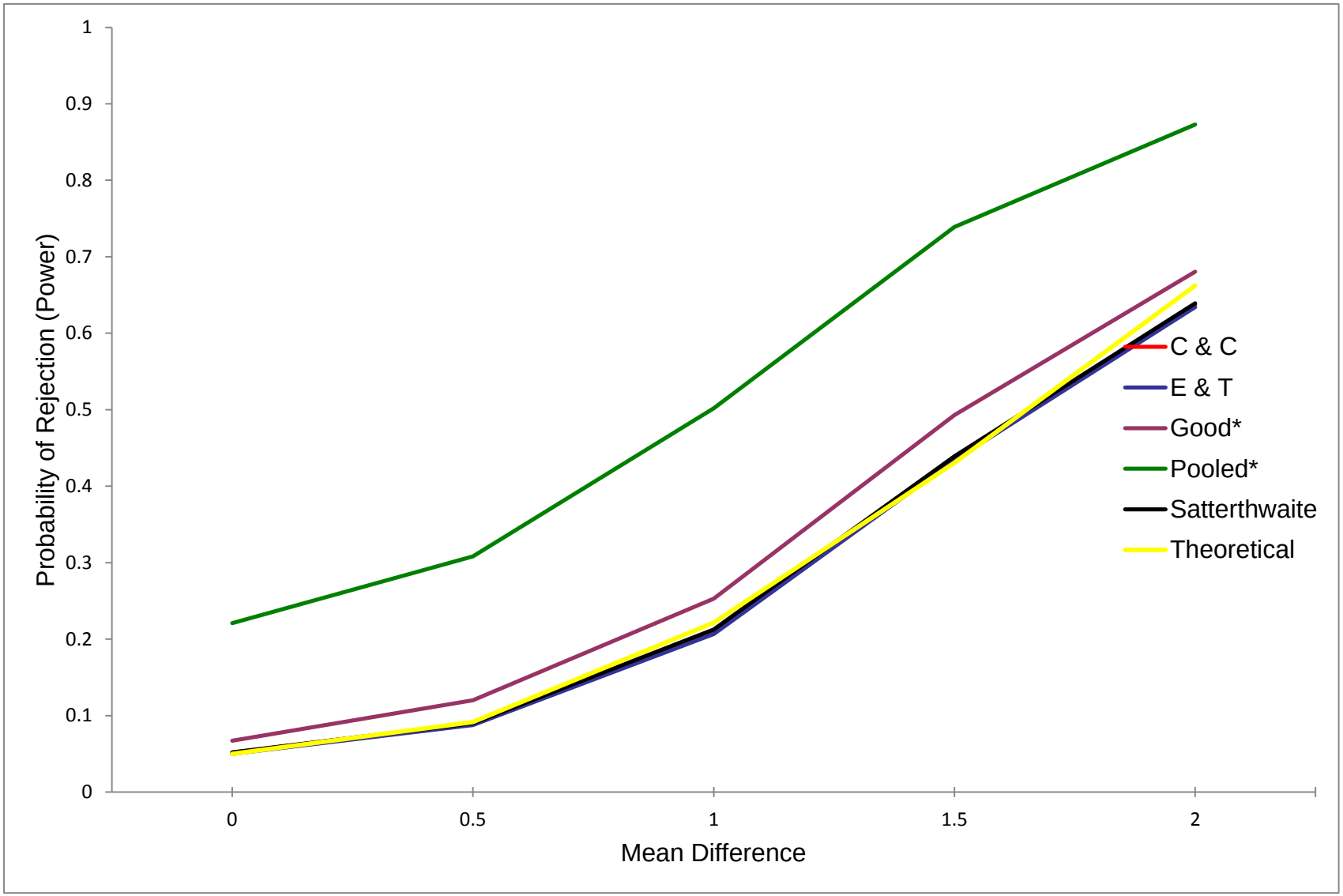

Figure B41. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

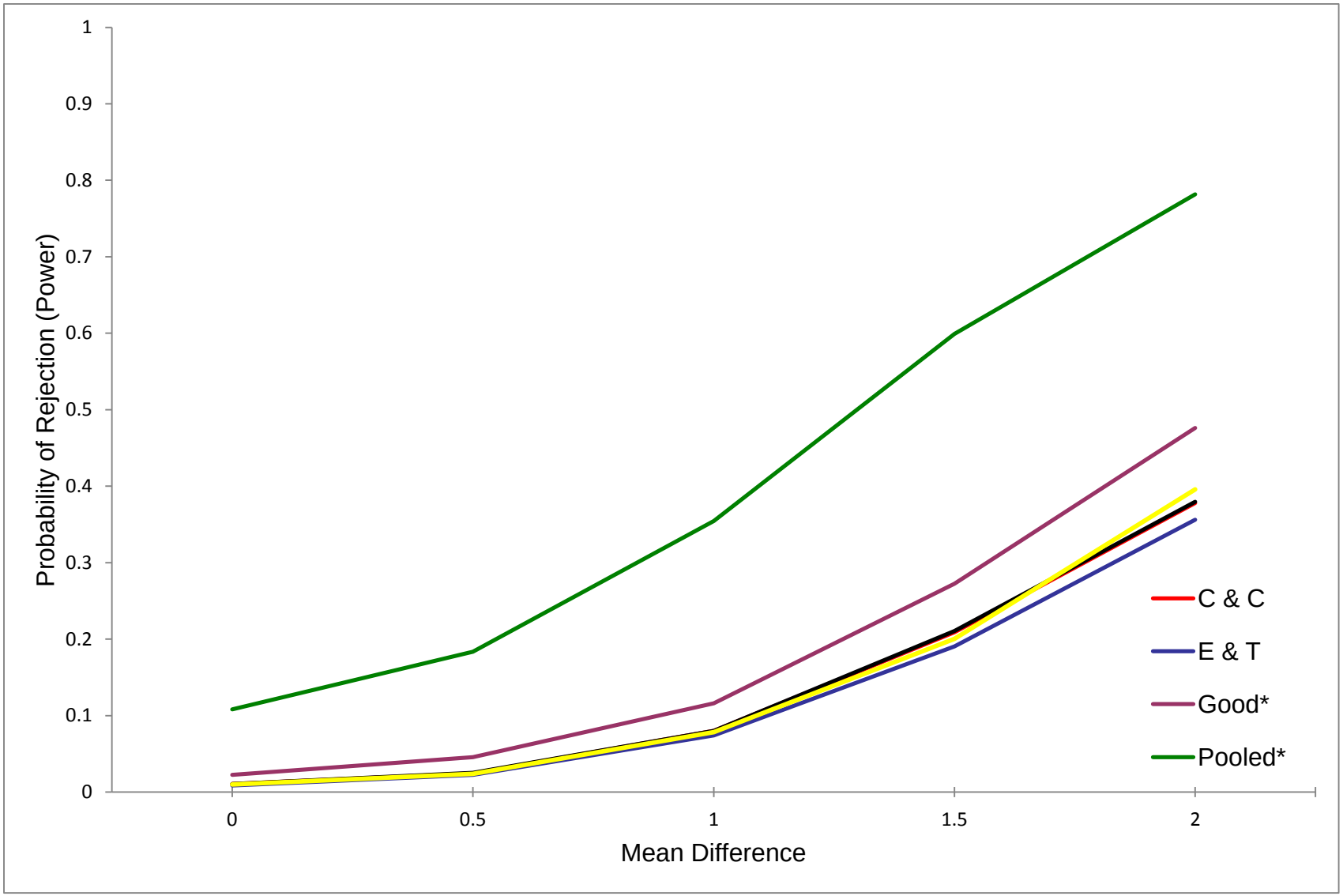

Figure B42. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 75$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 5.0 (i.e., $n_1 = 25$, $n_2 = 125$)

Table B28

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0360	0.0060
E & T	0.0430	0.0085
Good	0.0560	0.0145
Pooled	0.0005*	<.0005*
Satterthwaite	0.0435	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B29

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0345	0.0075
E & T	0.0435	0.0130
Good	0.0610	0.0210*
Pooled	0.0040*	0.0005*
Satterthwaite	0.0450	0.0125

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B30

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0435	0.0105
E & T	0.0505	0.0140
Good	0.0760*	0.0265*
Pooled	0.0155*	0.0010*
Satterthwaite	0.0530	0.0155

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B31

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0480	0.0100
E & T	0.0505	0.0105
Good	0.0885*	0.0290*
Pooled	0.0565	0.0125
Satterthwaite	0.0530	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B32

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0410	0.0085
E & T	0.0385	0.0080
Good	0.0840*	0.0280*
Pooled	0.1090*	0.0380*
Satterthwaite	0.0475	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B33

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0475	0.0085
E & T	0.0435	0.0080
Good	0.0980*	0.0385*
Pooled	0.2050*	0.1000*
Satterthwaite	0.0490	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B34

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0585	0.0140
E & T	0.0515	0.0090
Good	0.1190*	0.0475*
Pooled	0.3320*	0.2140*
Satterthwaite	0.0590	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table B35

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0525	0.0460	0.0500	0.0450	0.0425	0.0480	0.0450
	0.5	0.9985	0.9530	0.8270	0.5955	0.3880	0.2100	0.0885
	1.0	1.0000	1.0000	1.0000	0.9935	0.9070	0.6530	0.2180
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9415	0.4405
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980	0.6575
Efron & Tibshirani	0.0	0.0540	0.0510	0.0515	0.0485	0.0420	0.0480	0.0435
	0.5	0.9985	0.9540	0.8335	0.5975	0.3860	0.2075	0.0840
	1.0	1.0000	1.0000	1.0000	0.9935	0.9055	0.6490	0.2135
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9390	0.4320
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980	0.6465
Good	0.0	0.0570	0.0575	0.0650	0.0600	0.0605	0.0630	0.0650
	0.5	0.9985	0.9590	0.8470	0.6340	0.4390	0.2495	0.1170
	1.0	1.0000	1.0000	1.0000	0.9945	0.9325	0.7000	0.2555
	1.5	1.0000	1.0000	1.0000	1.0000	0.9985	0.9555	0.4910
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.7035
Pooled	0.0	<.0005	0.0020	0.0120	0.0470	0.1145	0.1955	0.3315
	0.5	0.8350	0.7575	0.6880	0.6220	0.5640	0.4860	0.4075
	1.0	1.0000	1.0000	1.0000	0.9955	0.9685	0.8830	0.5990
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9895	0.8005
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9310
Satterthwaite	0.0	0.0550	0.0505	0.0540	0.0490	0.0440	0.0495	0.0455
	0.5	0.9985	0.9555	0.8345	0.6020	0.3925	0.2120	0.0890
	1.0	1.0000	1.0000	1.0000	0.9935	0.9100	0.6550	0.2180
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9420	0.4410
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.6580

Table B36

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 25$, $n_2 = 125$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0095	0.0075	0.0075	0.0065	0.0075	0.0095	0.0085
	0.5	0.9880	0.8480	0.6000	0.3265	0.1600	0.0720	0.0220
	1.0	1.0000	1.0000	0.9995	0.9525	0.7165	0.3875	0.0800
	1.5	1.0000	1.0000	1.0000	1.0000	0.9865	0.7995	0.2155
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9780	0.3965
Efron & Tibshirani	0.0	0.0110	0.0080	0.0085	0.0070	0.0070	0.0090	0.0080
	0.5	0.9890	0.8580	0.6160	0.3330	0.1575	0.0680	0.0205
	1.0	1.0000	1.0000	0.9995	0.9545	0.7065	0.3805	0.0690
	1.5	1.0000	1.0000	1.0000	1.0000	0.9850	0.7855	0.2020
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9745	0.3735
Good	0.0	0.0125	0.0125	0.0170	0.0130	0.0155	0.0175	0.0150
	0.5	0.9915	0.8780	0.6735	0.4115	0.2245	0.1090	0.0360
	1.0	1.0000	1.0000	1.0000	0.9765	0.7945	0.4855	0.1160
	1.5	1.0000	1.0000	1.0000	1.0000	0.9900	0.8680	0.2850
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9895	0.4900
Pooled	0.0	<.0005	<.0005	0.0015	0.0085	0.0375	0.0875	0.2000
	0.5	0.4005	0.3965	0.3925	0.3620	0.3650	0.3195	0.2800
	1.0	1.0000	0.9990	0.9960	0.9720	0.9080	0.7755	0.4640
	1.5	1.0000	1.0000	1.0000	1.0000	0.9985	0.9765	0.7080
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8800
Satterthwaite	0.0	0.0110	0.0080	0.0090	0.0075	0.0075	0.0095	0.0085
	0.5	0.9885	0.8630	0.6225	0.3435	0.1680	0.0735	0.0225
	1.0	1.0000	1.0000	0.9995	0.9580	0.7250	0.3935	0.0805
	1.5	1.0000	1.0000	1.0000	1.0000	0.9865	0.8050	0.2165
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9780	0.3975

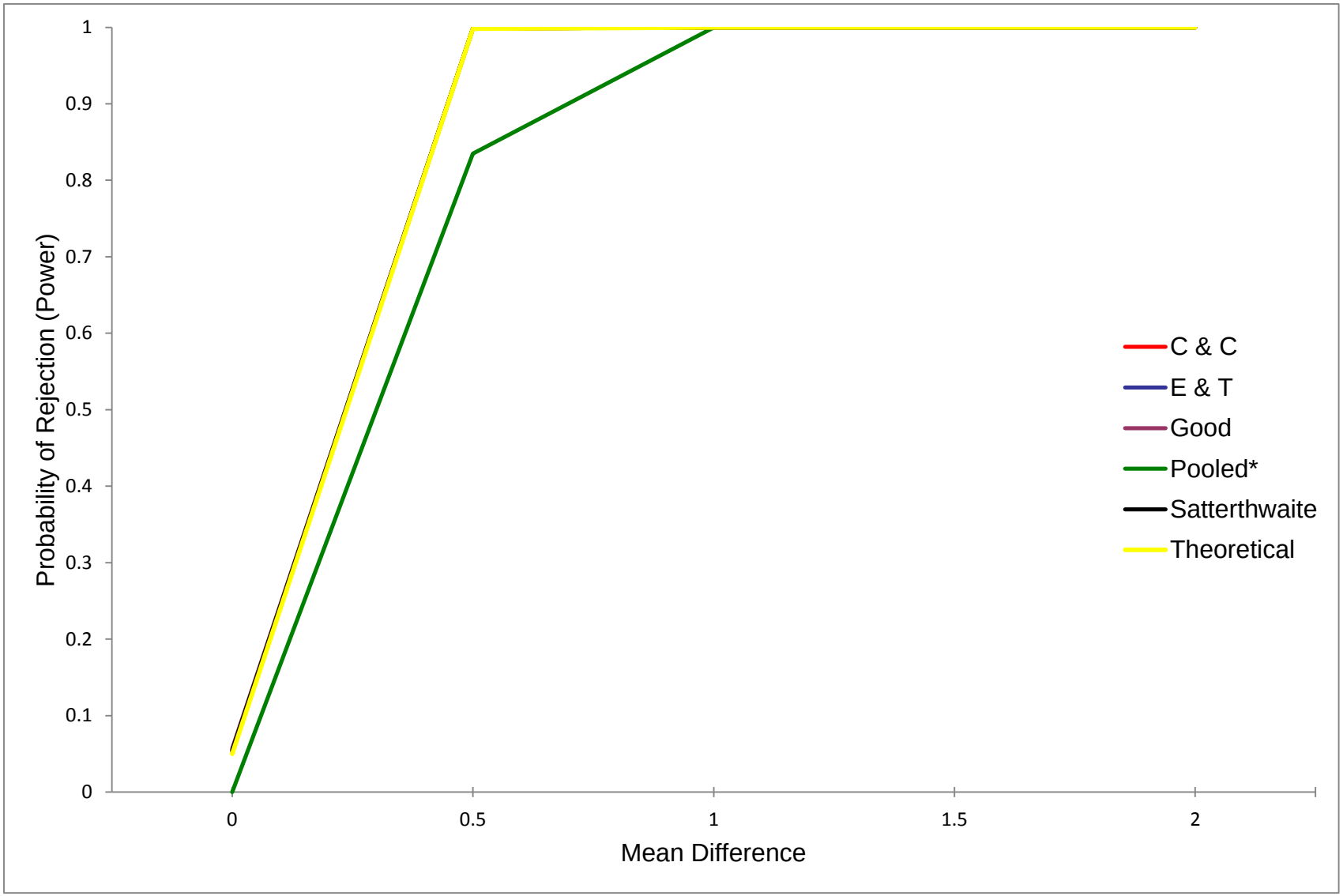

Figure B43. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

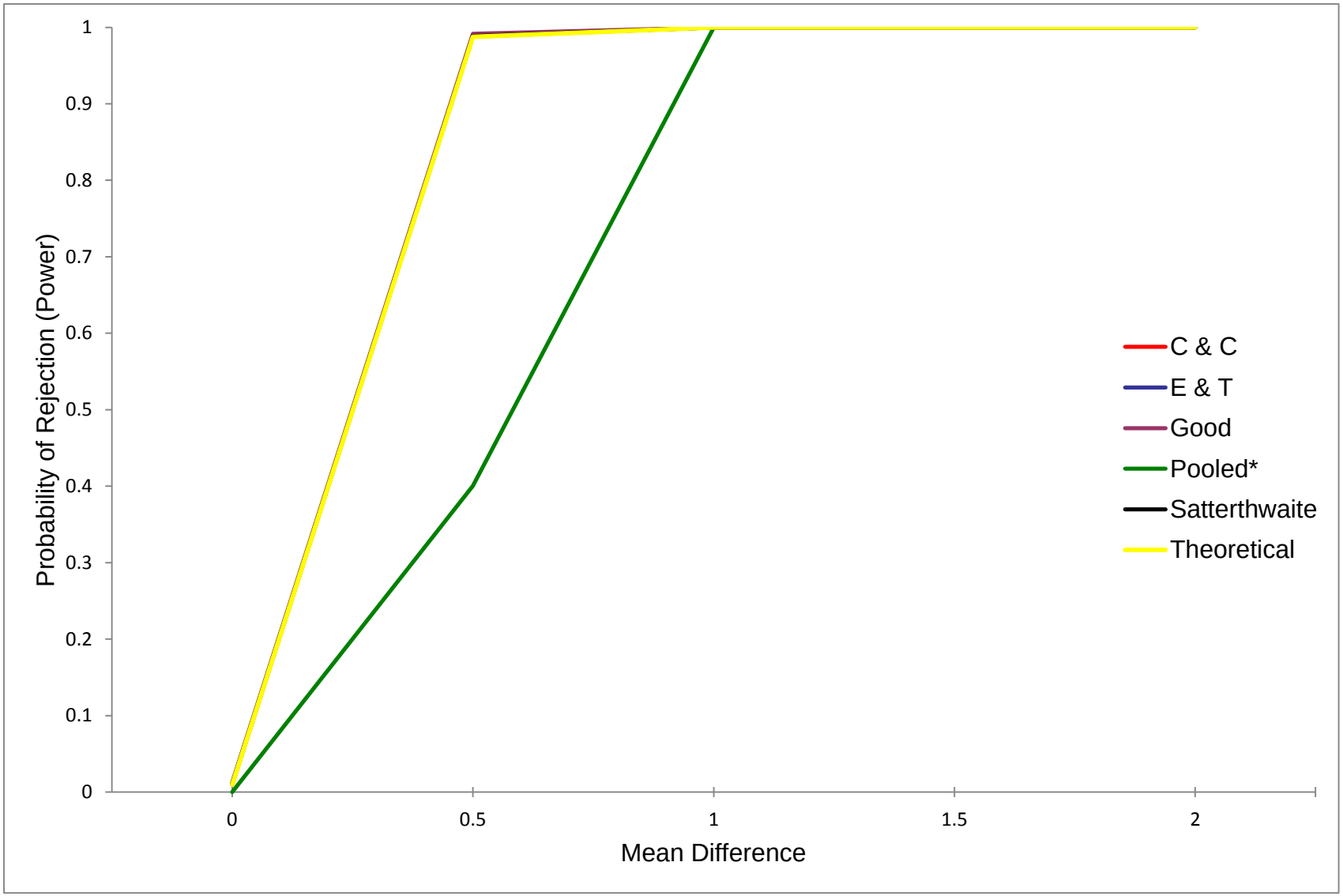

Figure B44. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

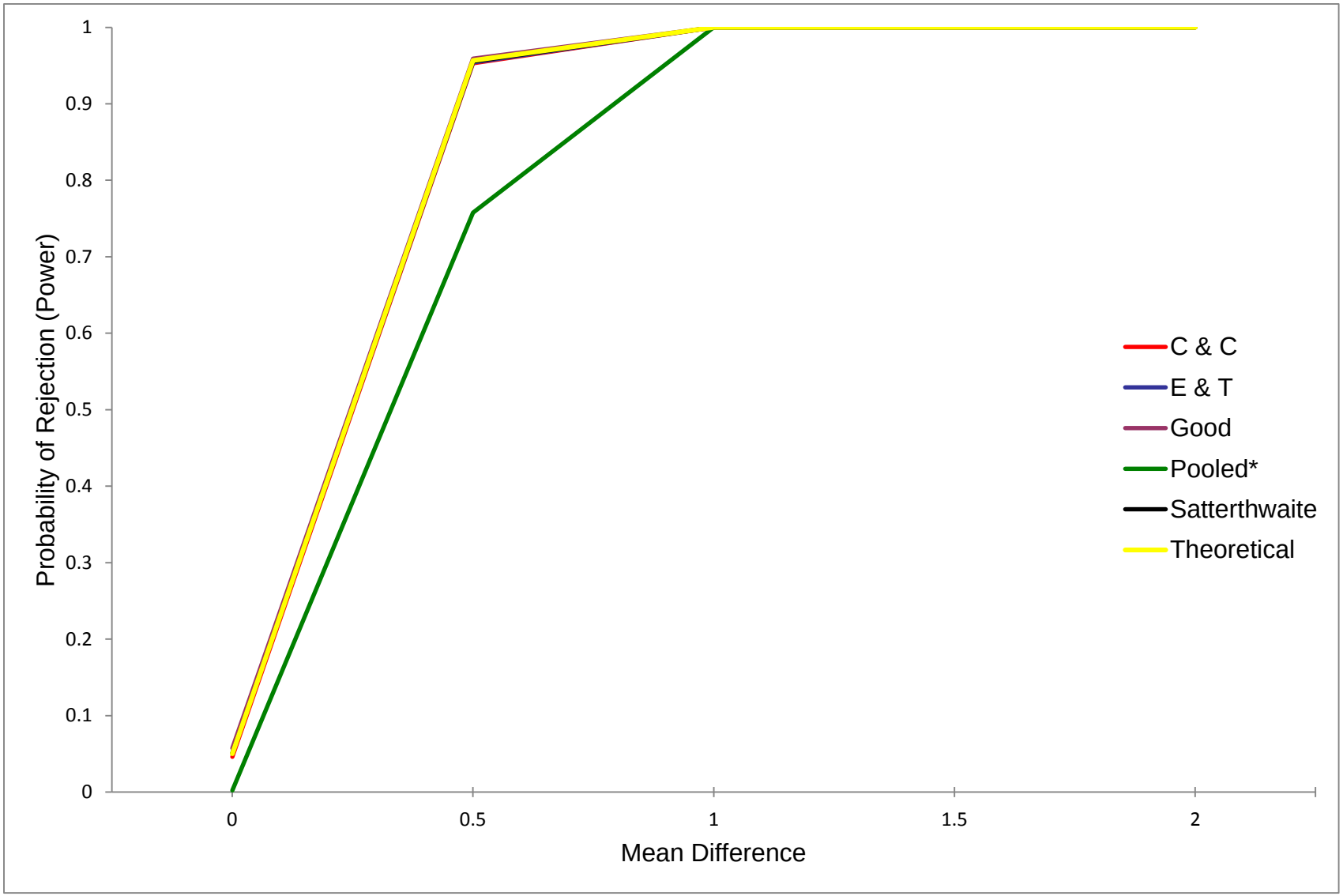

Figure B45. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

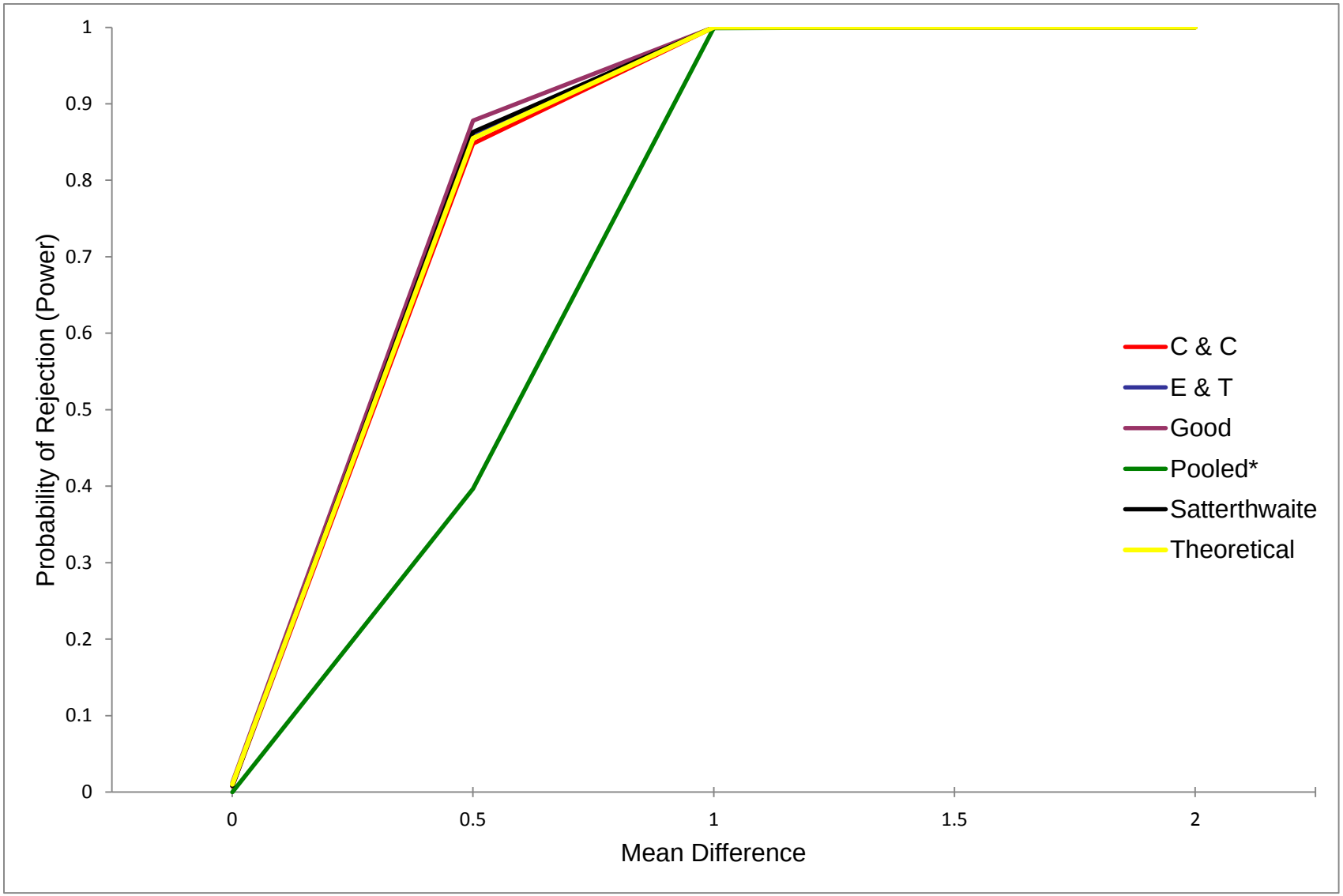

Figure B46. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

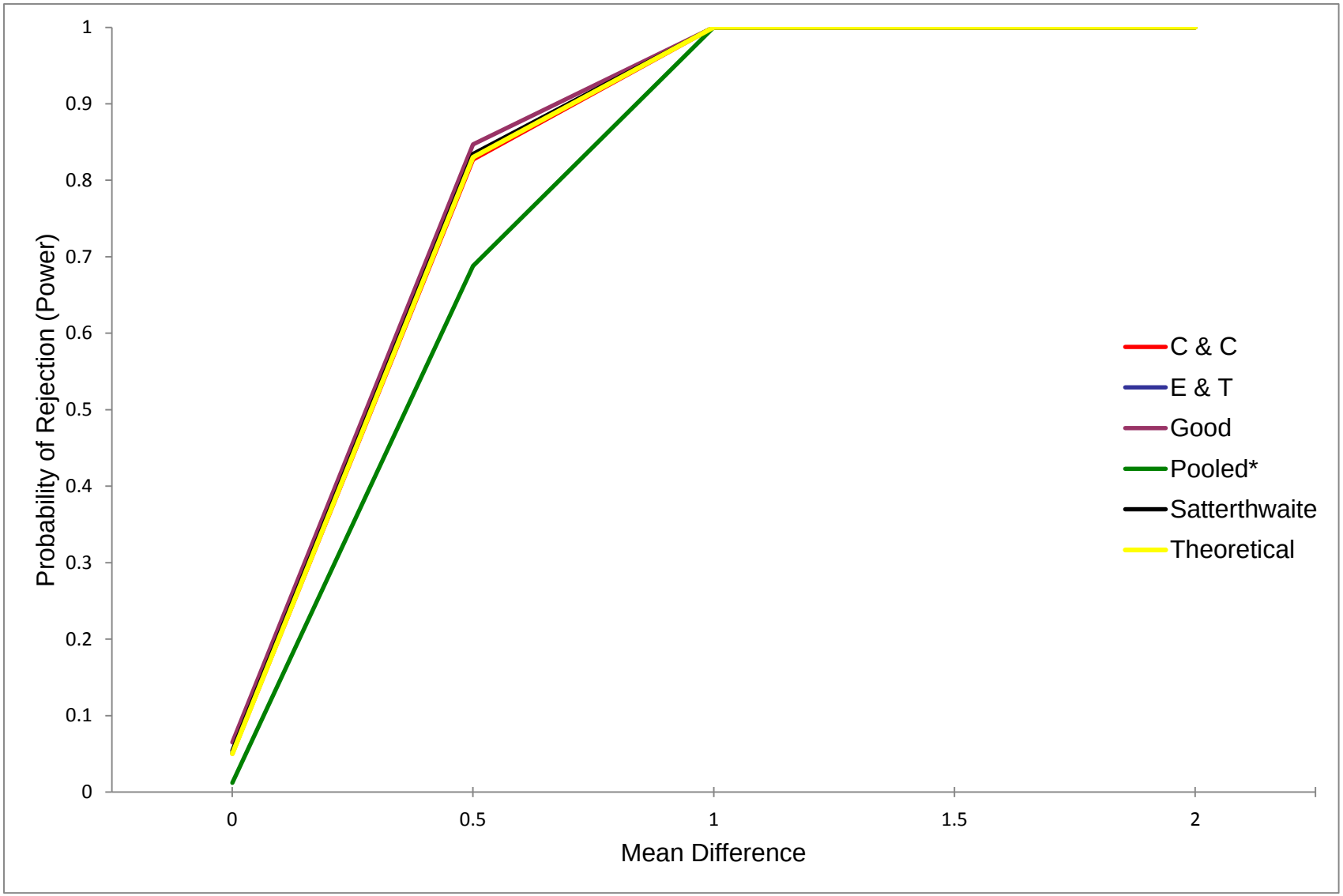

Figure B47. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

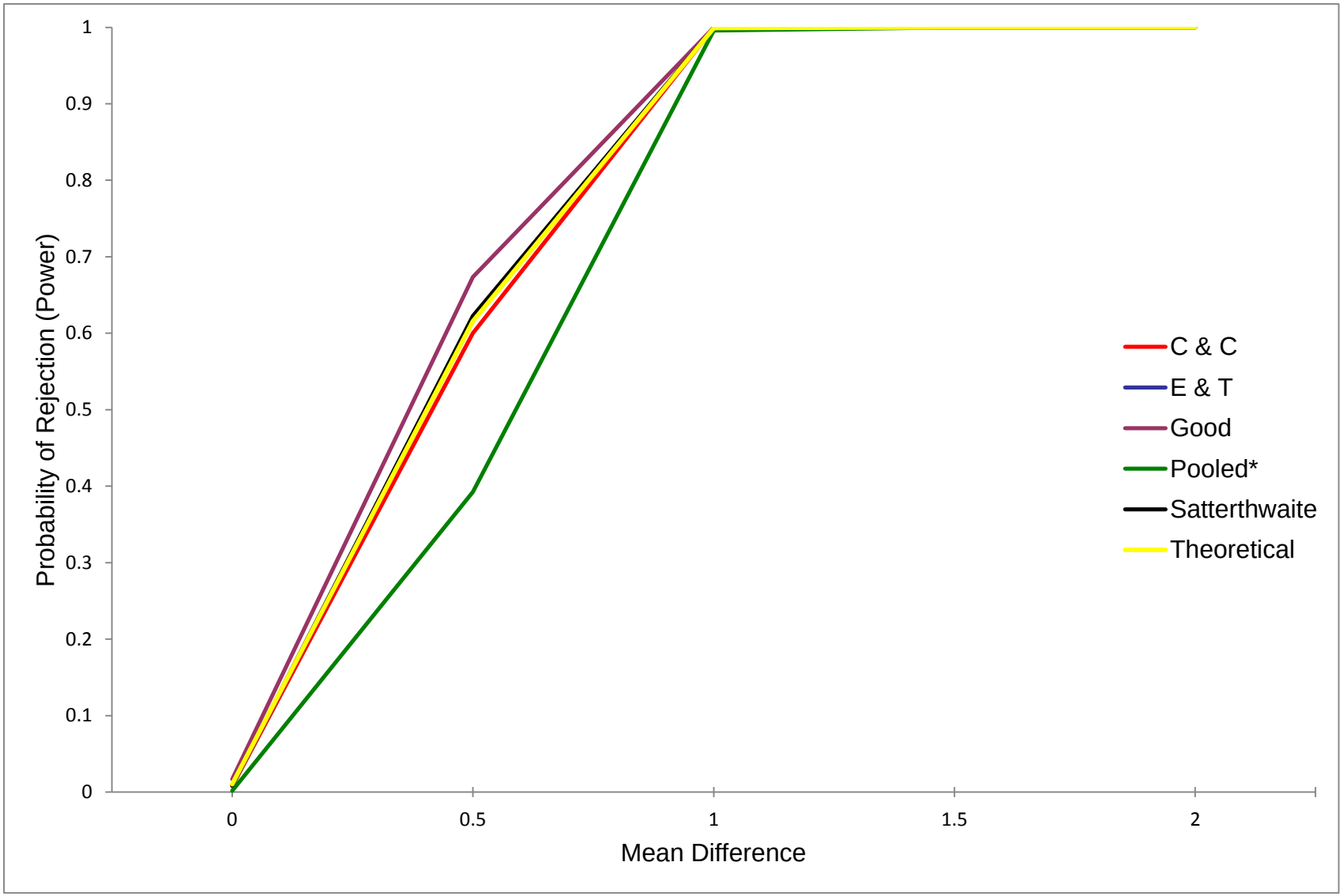

Figure B48. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

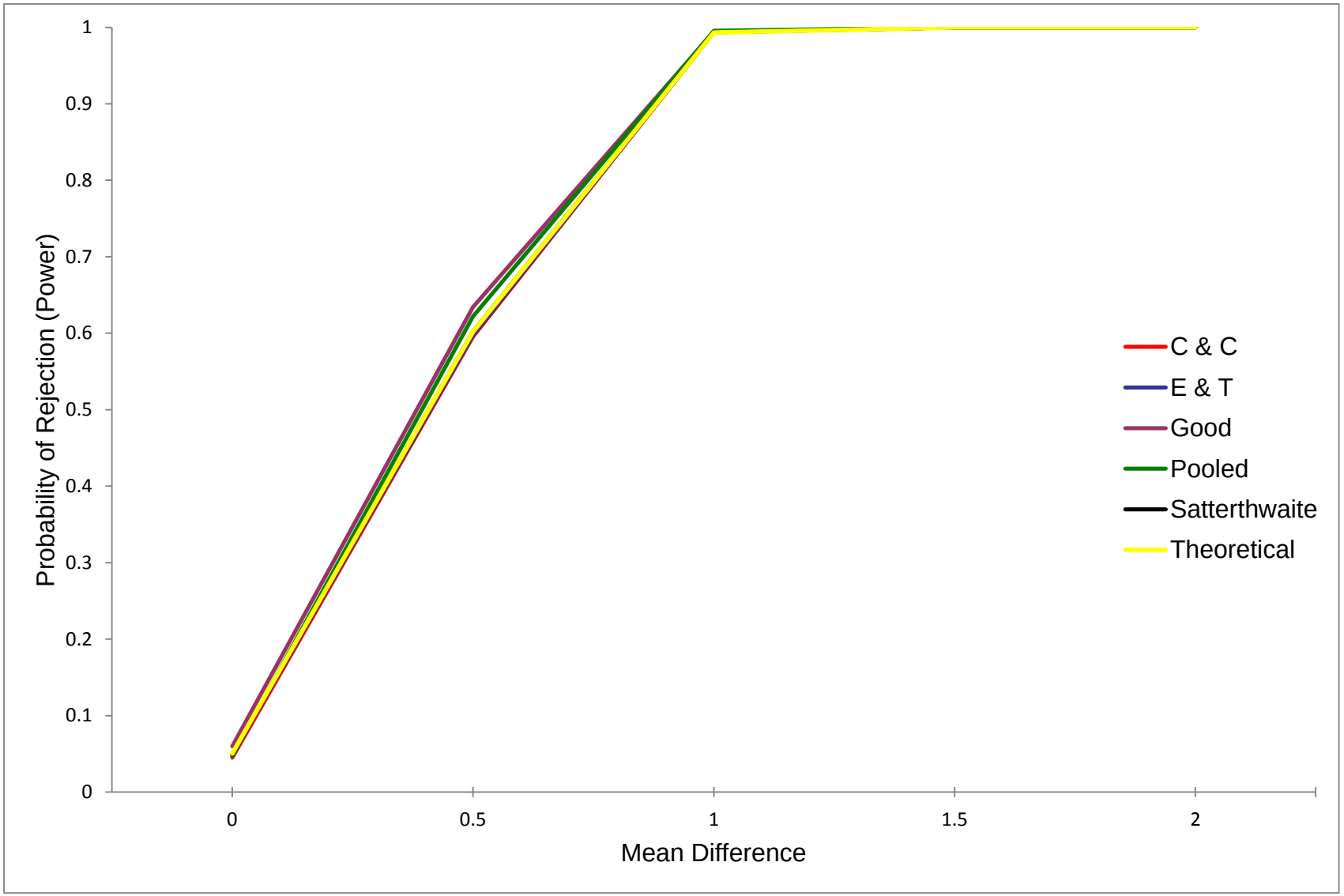

Figure B49. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

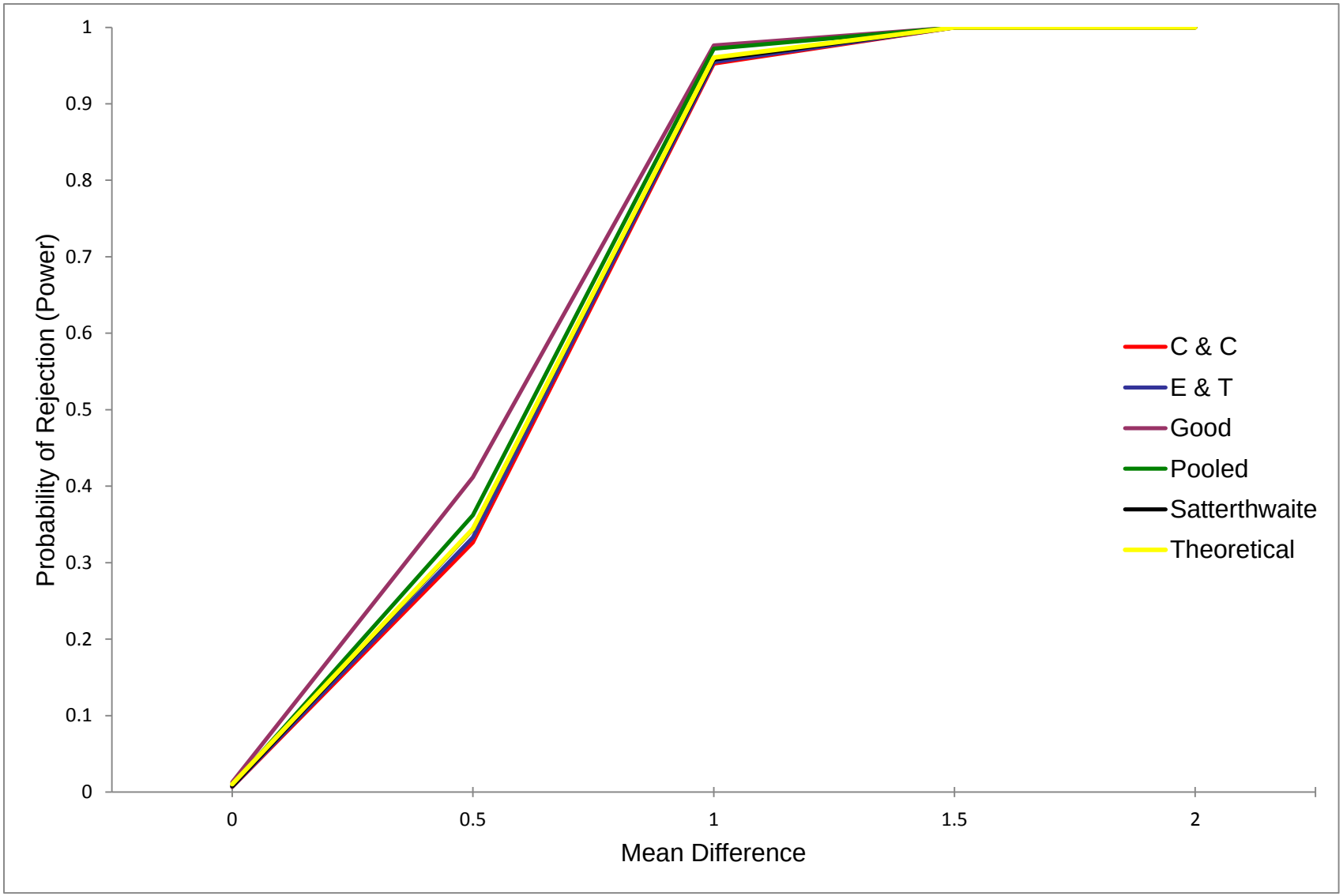

Figure B50. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

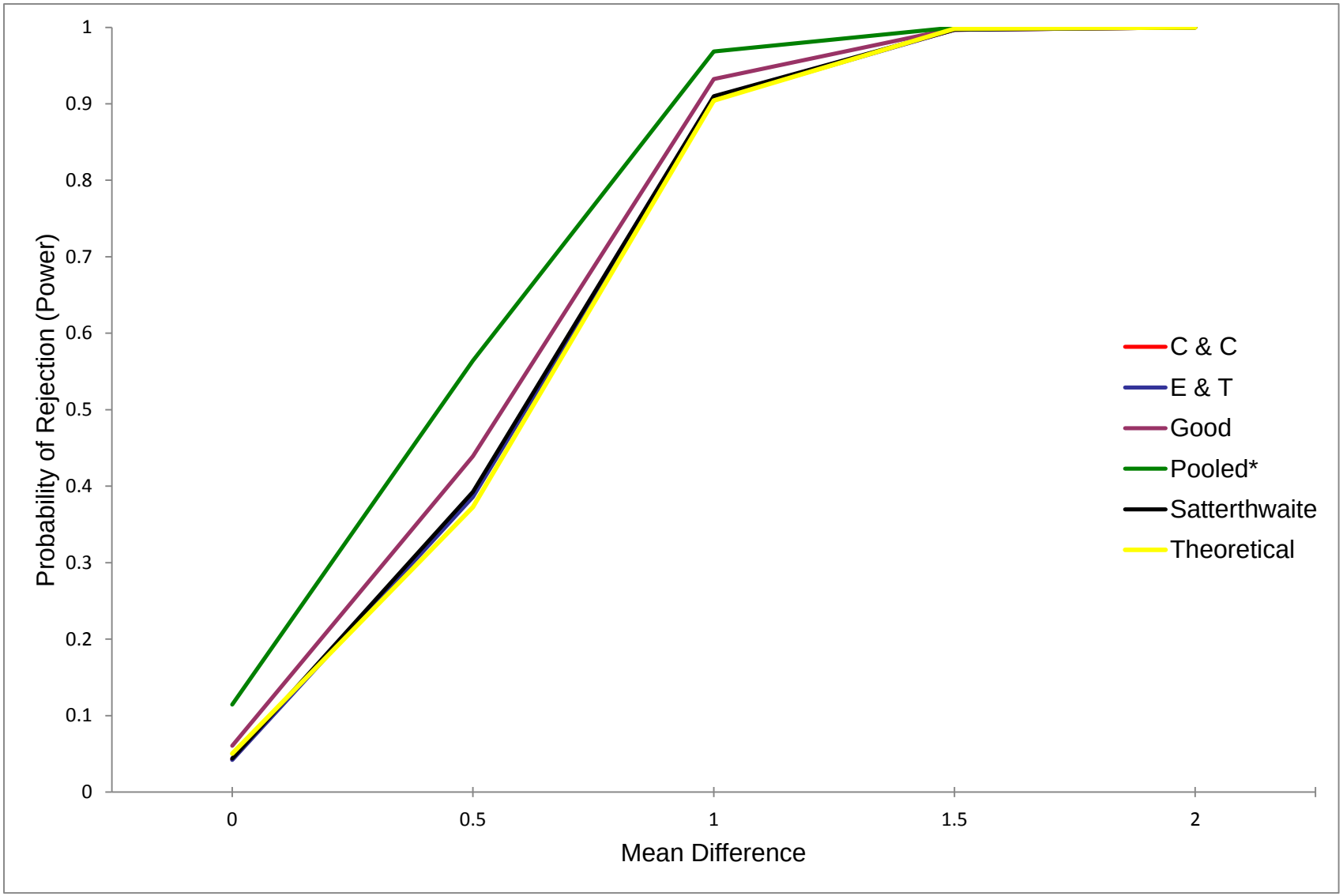

Figure B51. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

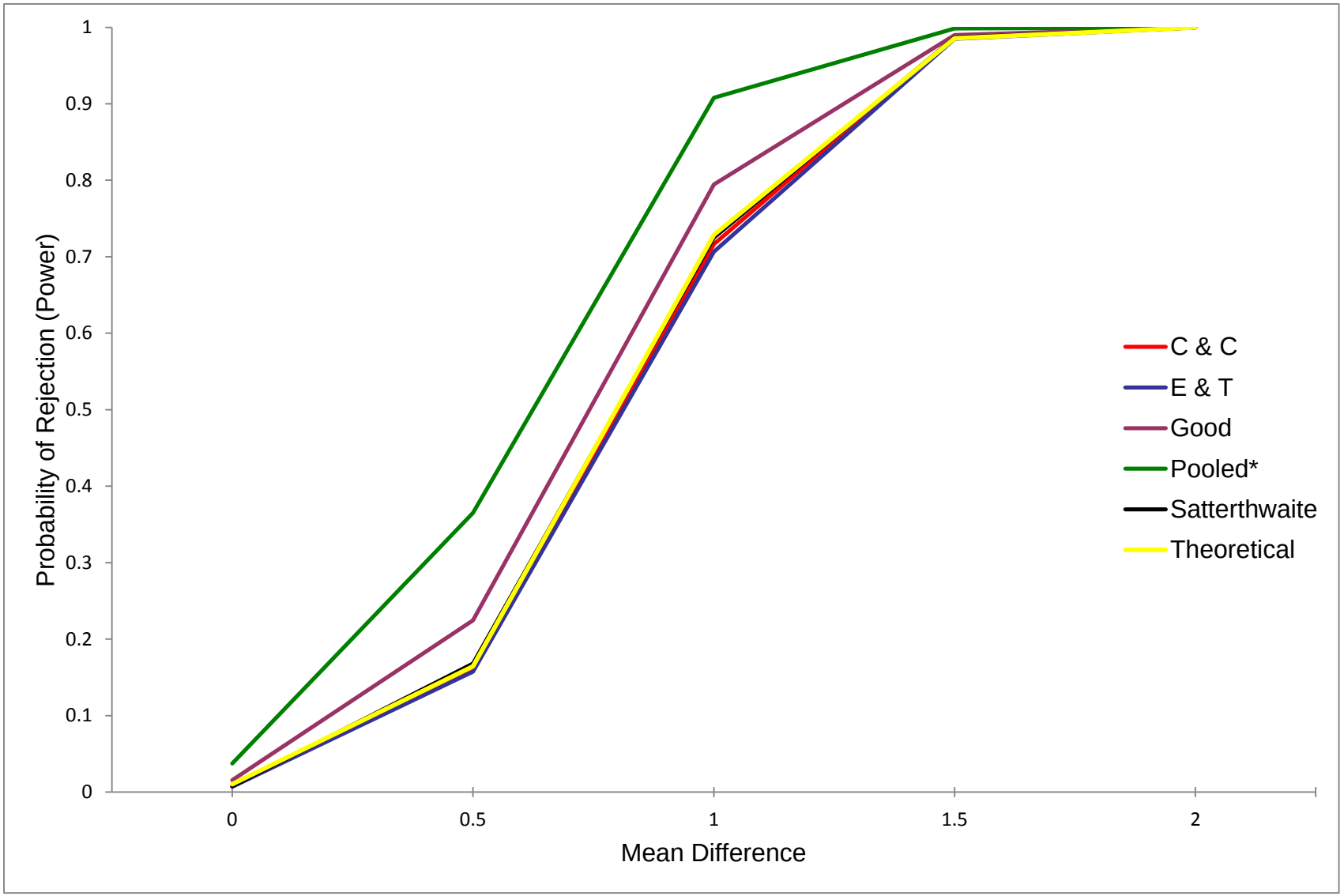

Figure B52. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

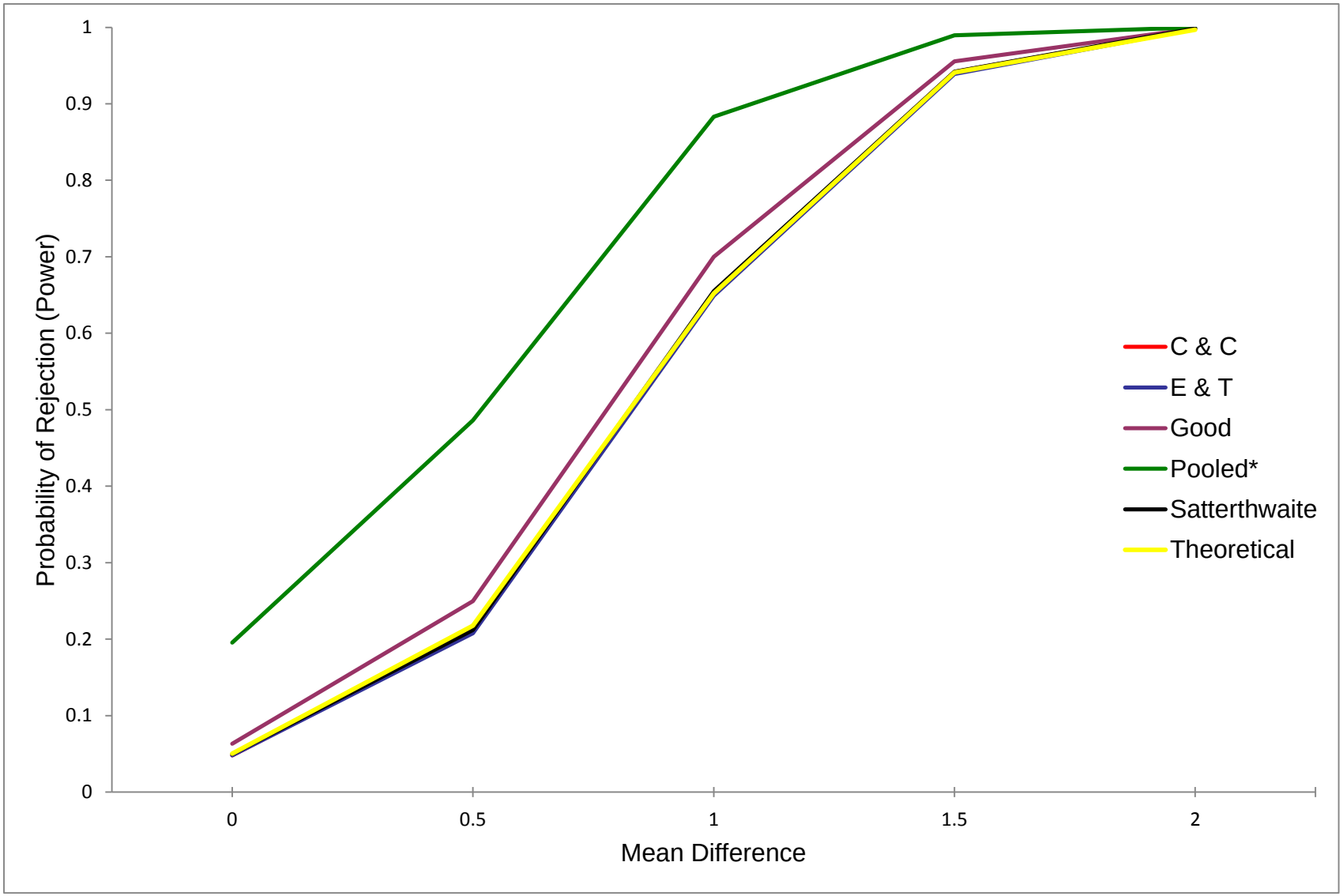

Figure B53. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

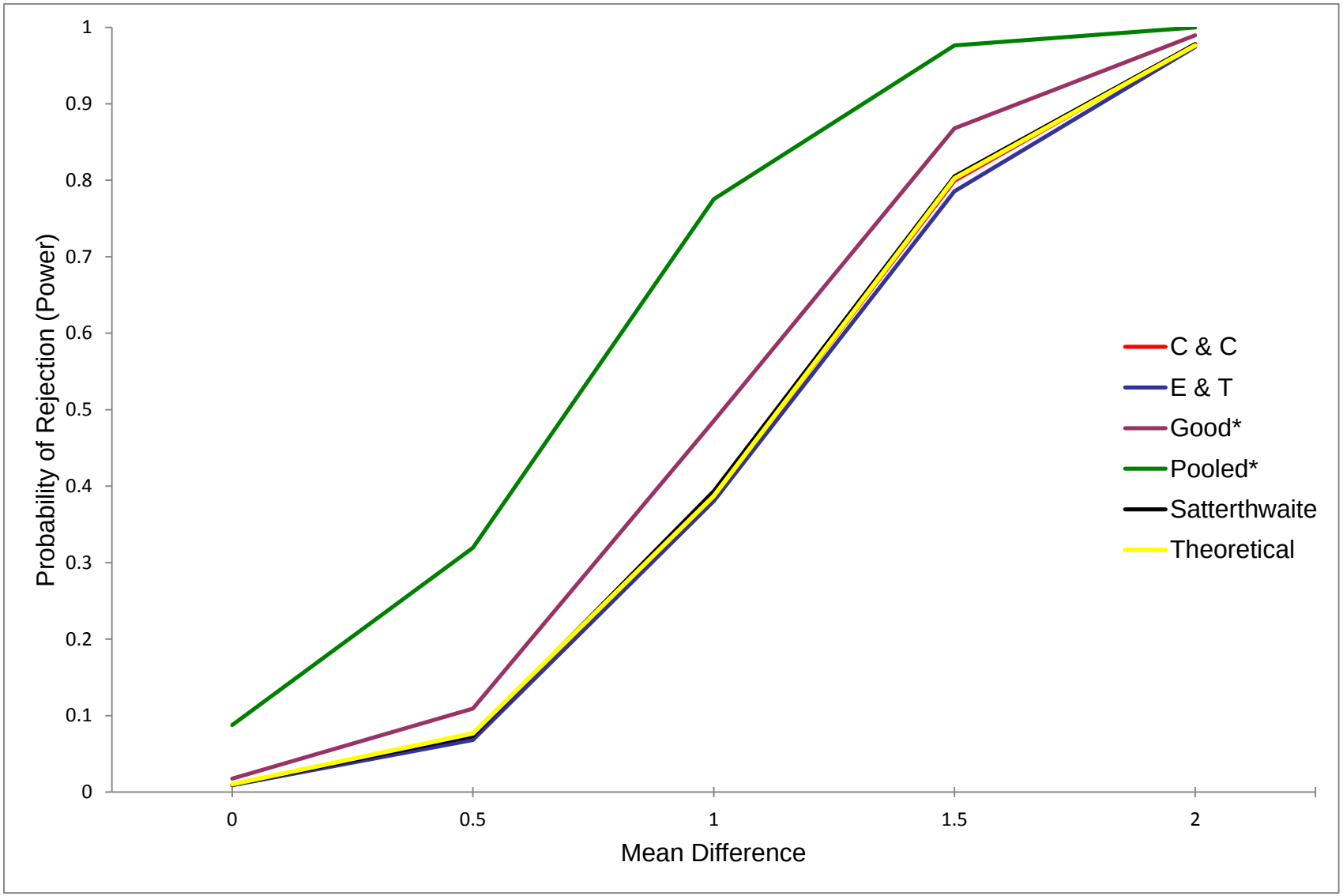

Figure B54. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

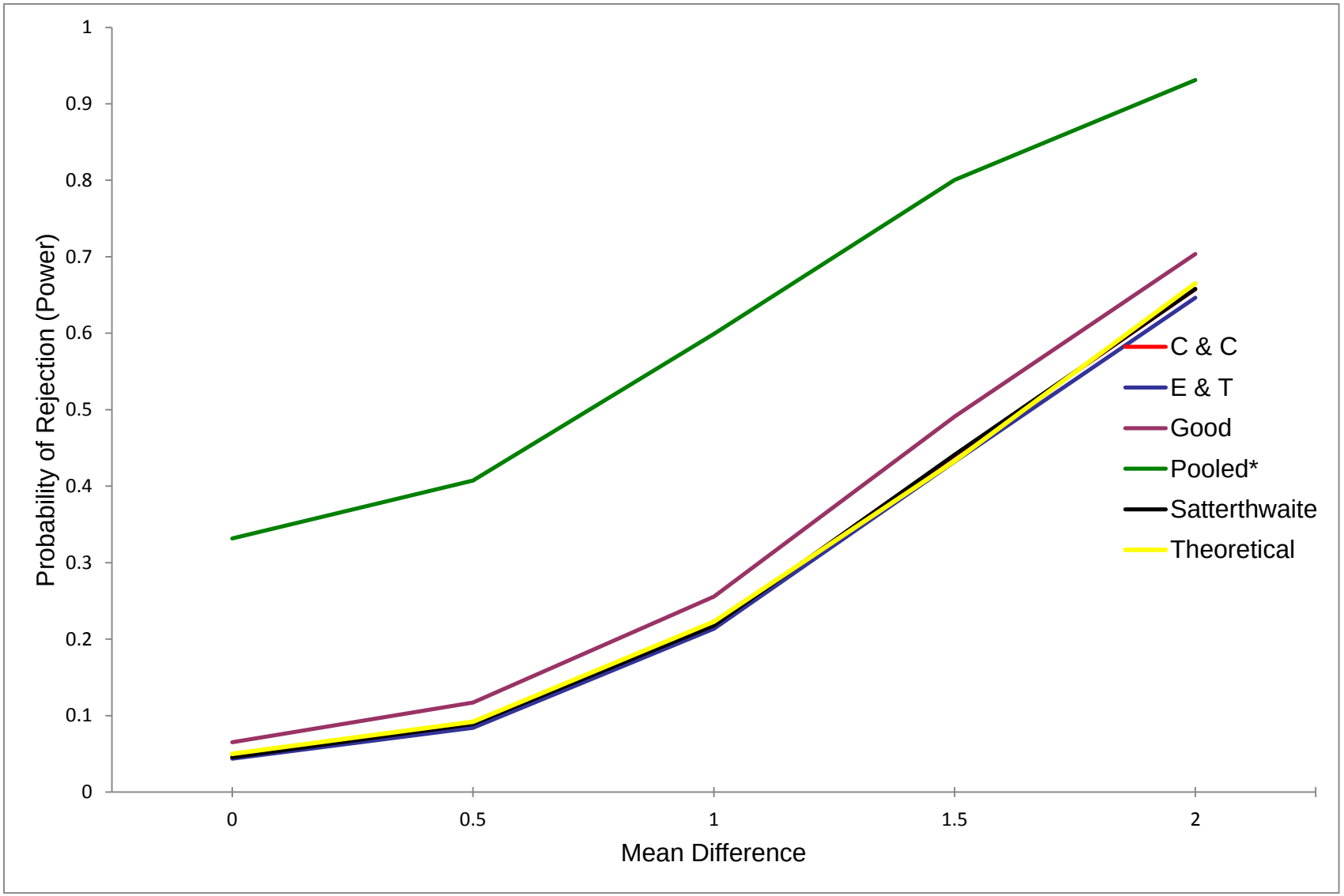

Figure B55. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

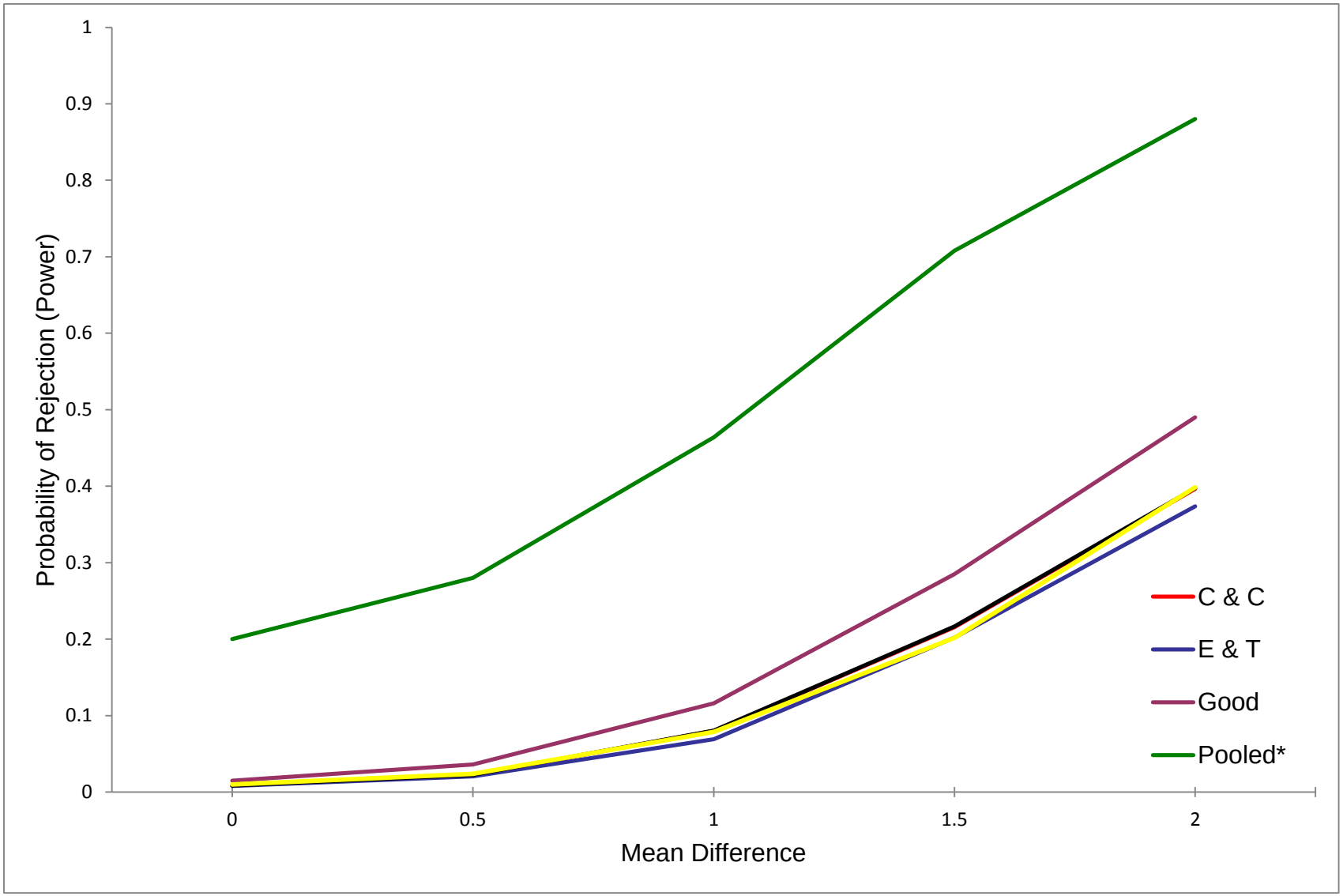

Figure B56. Power curves for unequal group sample sizes when $n_1 = 25$, $n_2 = 125$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

APPENDIX C: TYPE I ERROR RATE TABLES, POWER TABLES, AND POWER CURVES, WHEN THE SAMPLE SIZE OF GROUP 1 (n_1) WAS 40

Sample-size Ratio was 1.0 (i.e., Equal Sample Size or $n_1 = n_2 = 40$)

Table C1

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0430	0.0080
E & T	0.0430	0.0080
Good	0.0555	0.0120
Pooled	0.0460	0.0105
Satterthwaite	0.0440	0.0085

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C2

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0485	0.0090
E & T	0.0505	0.0095
Good	0.0615	0.0130
Pooled	0.0530	0.0110
Satterthwaite	0.0525	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C3

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0465	0.0065
E & T	0.0505	0.0080
Good	0.0570	0.0110
Pooled	0.0505	0.0075
Satterthwaite	0.0505	0.0075

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C4

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0505	0.0070
E & T	0.0550	0.0090
Good	0.0620	0.0125
Pooled	0.0545	0.0095
Satterthwaite	0.0545	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C5

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0450	0.0085
E & T	0.0465	0.0115
Good	0.0530	0.0150
Pooled	0.0480	0.0110
Satterthwaite	0.0475	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C6

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0435	0.0085
E & T	0.0455	0.0095
Good	0.0560	0.0130
Pooled	0.0480	0.0105
Satterthwaite	0.0455	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C7

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0505	0.0110
E & T	0.0505	0.0105
Good	0.0575	0.0165
Pooled	0.0535	0.0115
Satterthwaite	0.0510	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C8

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0430	0.0485	0.0465	0.0505	0.0450	0.0435	0.0505
	0.5	0.8580	0.7895	0.7135	0.5940	0.4190	0.2530	0.1180
	1.0	1.0000	0.9990	1.0000	0.9920	0.9455	0.7780	0.3465
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9825	0.6105
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8480
Efron & Tibshirani	0.0	0.0430	0.0505	0.0505	0.0550	0.0465	0.0455	0.0505
	0.5	0.8570	0.7925	0.7220	0.6025	0.4280	0.2555	0.1190
	1.0	1.0000	0.9990	1.0000	0.9925	0.9480	0.7840	0.3425
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9830	0.6120
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8505
Good	0.0	0.0555	0.0615	0.0570	0.0620	0.0530	0.0560	0.0575
	0.5	0.8770	0.8125	0.7485	0.6220	0.4470	0.2775	0.1355
	1.0	1.0000	0.9990	1.0000	0.9950	0.9545	0.8050	0.3700
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9870	0.6465
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8665
Pooled	0.0	0.0460	0.0530	0.0505	0.0545	0.0480	0.0480	0.0535
	0.5	0.8625	0.7960	0.7250	0.6040	0.4310	0.2590	0.1255
	1.0	1.0000	0.9990	1.0000	0.9925	0.9485	0.7885	0.3530
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9840	0.6250
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8590
Satterthwaite	0.0	0.0440	0.0525	0.0505	0.0545	0.0475	0.0455	0.0510
	0.5	0.8590	0.7945	0.7235	0.6040	0.4300	0.2570	0.1205
	1.0	1.0000	0.9990	1.0000	0.9925	0.9480	0.7840	0.3490
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9840	0.6130
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8510

Table C9

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = n_2 = 40$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0080	0.0090	0.0065	0.0070	0.0085	0.0085	0.0110
	0.5	0.6475	0.5420	0.4500	0.3245	0.1850	0.0940	0.0340
	1.0	1.0000	0.9980	0.9950	0.9595	0.8325	0.5460	0.1415
	1.5	1.0000	1.0000	1.0000	1.0000	0.9960	0.9390	0.3625
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.6465
Efron & Tibshirani	0.0	0.0080	0.0095	0.0080	0.0090	0.0115	0.0095	0.0105
	0.5	0.6445	0.5530	0.4735	0.3460	0.1995	0.0995	0.0335
	1.0	1.0000	0.9975	0.9965	0.9630	0.8445	0.5595	0.1430
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9415	0.3580
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9990	0.6435
Good	0.0	0.0120	0.0130	0.0110	0.0125	0.0150	0.0130	0.0165
	0.5	0.7095	0.6140	0.5165	0.3895	0.2270	0.1230	0.0450
	1.0	1.0000	0.9985	0.9975	0.9715	0.8730	0.6155	0.1780
	1.5	1.0000	1.0000	1.0000	1.0000	0.9975	0.9520	0.4165
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.7055
Pooled	0.0	0.0105	0.0110	0.0075	0.0095	0.0110	0.0105	0.0115
	0.5	0.6820	0.5765	0.4810	0.3525	0.2040	0.1040	0.0370
	1.0	1.0000	0.9985	0.9960	0.9635	0.8495	0.5740	0.1575
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9445	0.3830
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6680
Satterthwaite	0.0	0.0085	0.0100	0.0075	0.0095	0.0110	0.0105	0.0110
	0.5	0.6550	0.5635	0.4785	0.3525	0.2015	0.1025	0.0350
	1.0	1.0000	0.9980	0.9960	0.9635	0.8455	0.5645	0.1470
	1.5	1.0000	1.0000	1.0000	1.0000	0.9970	0.9430	0.3680
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6535

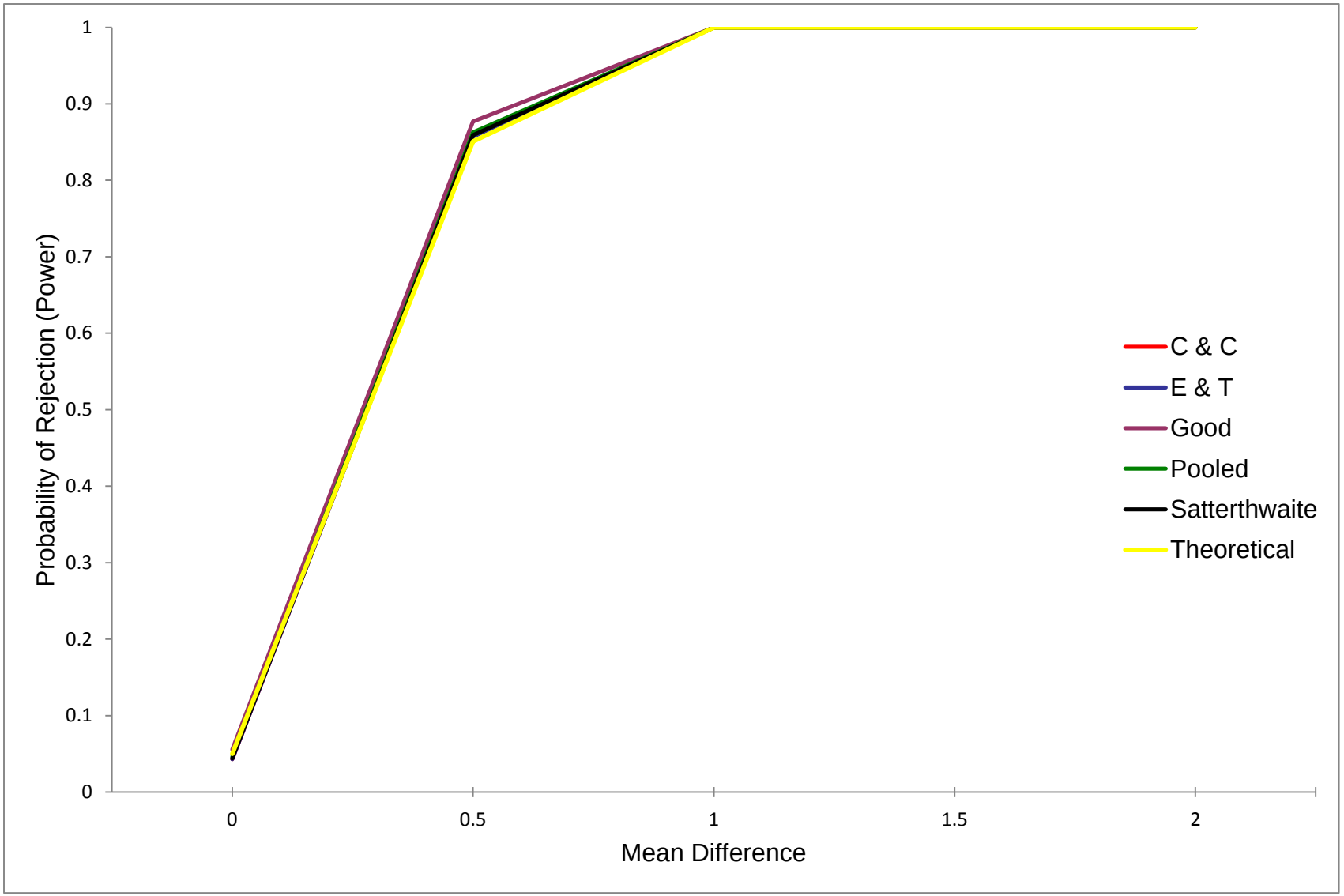

Figure C1. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

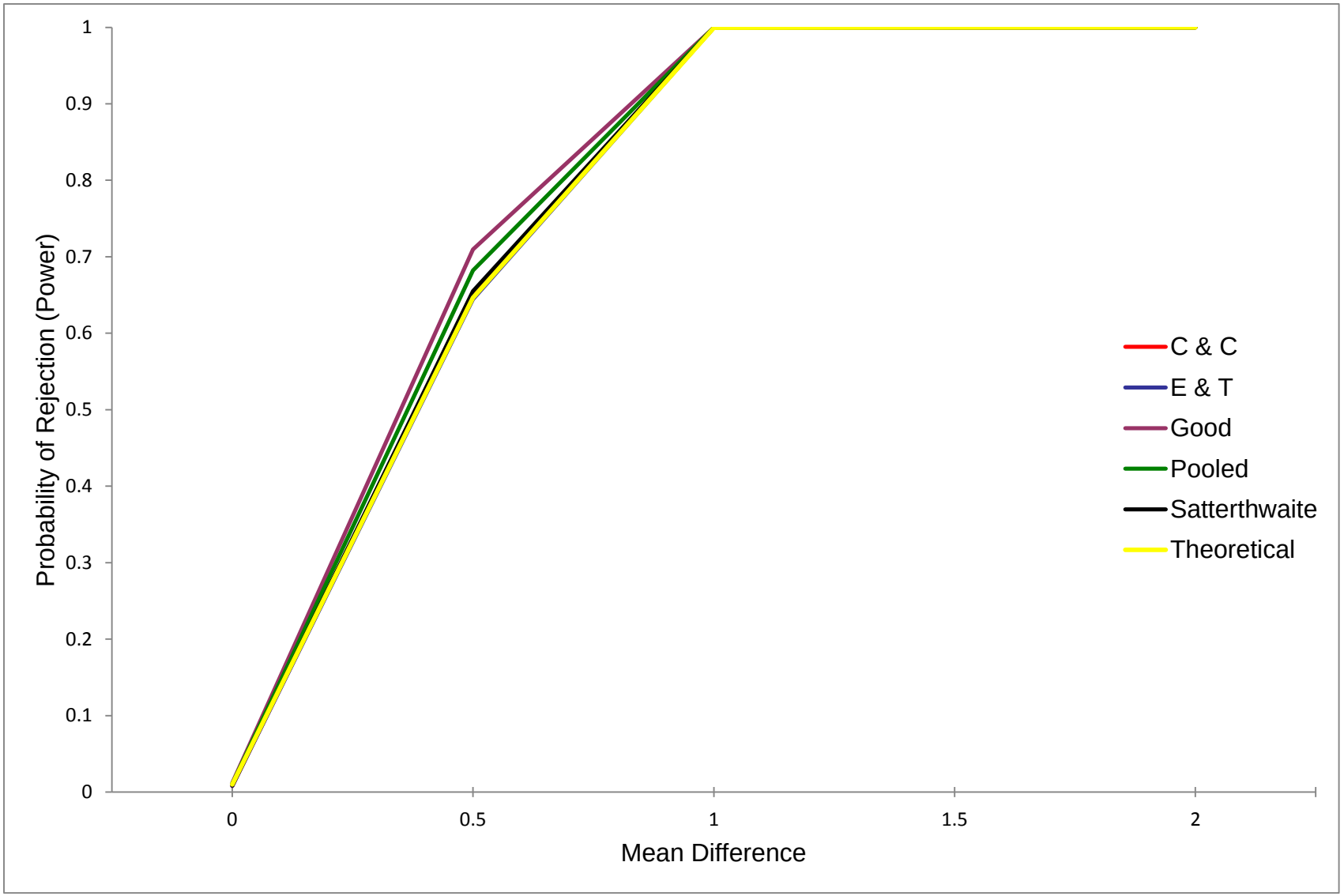

Figure C2. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

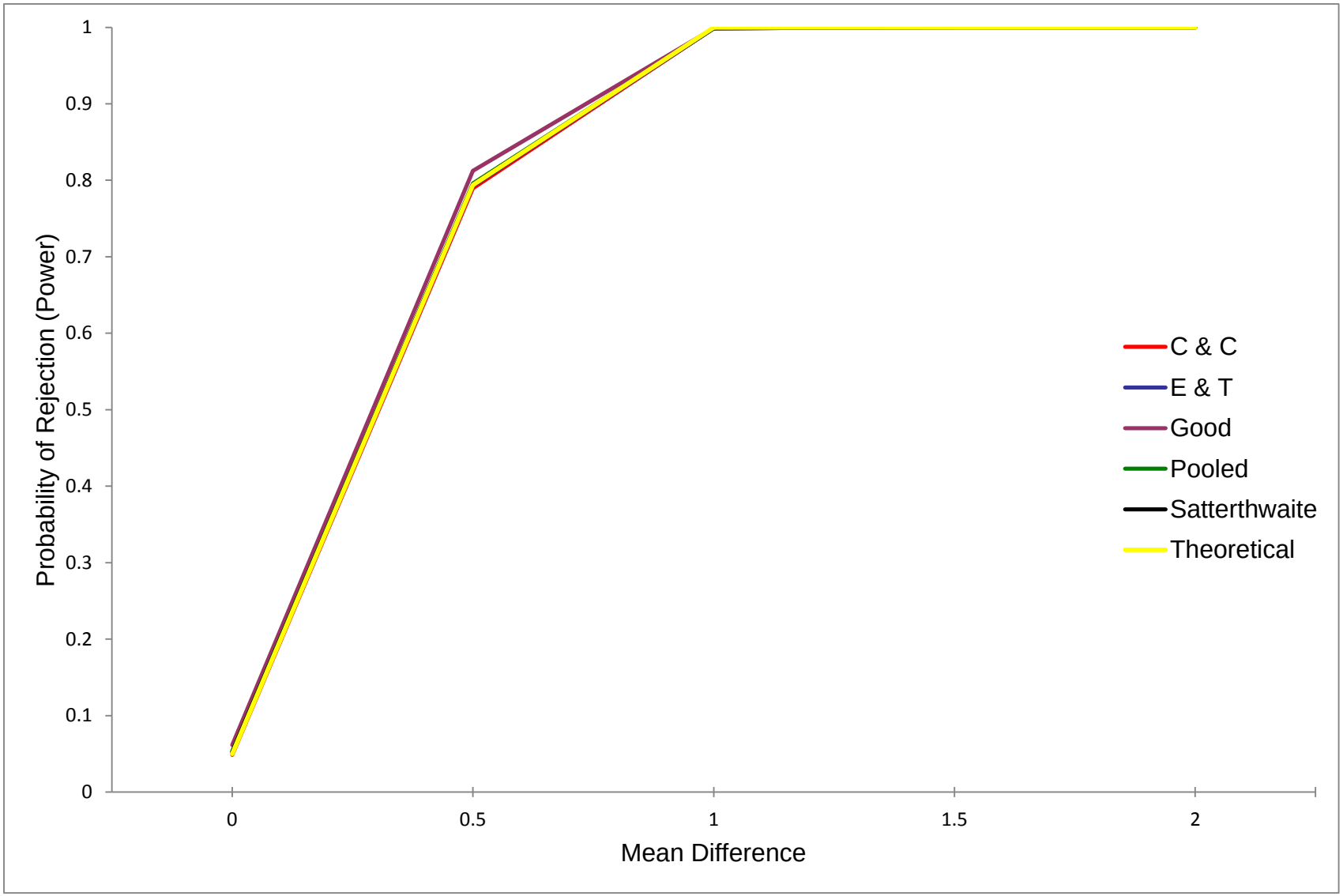

Figure C3. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

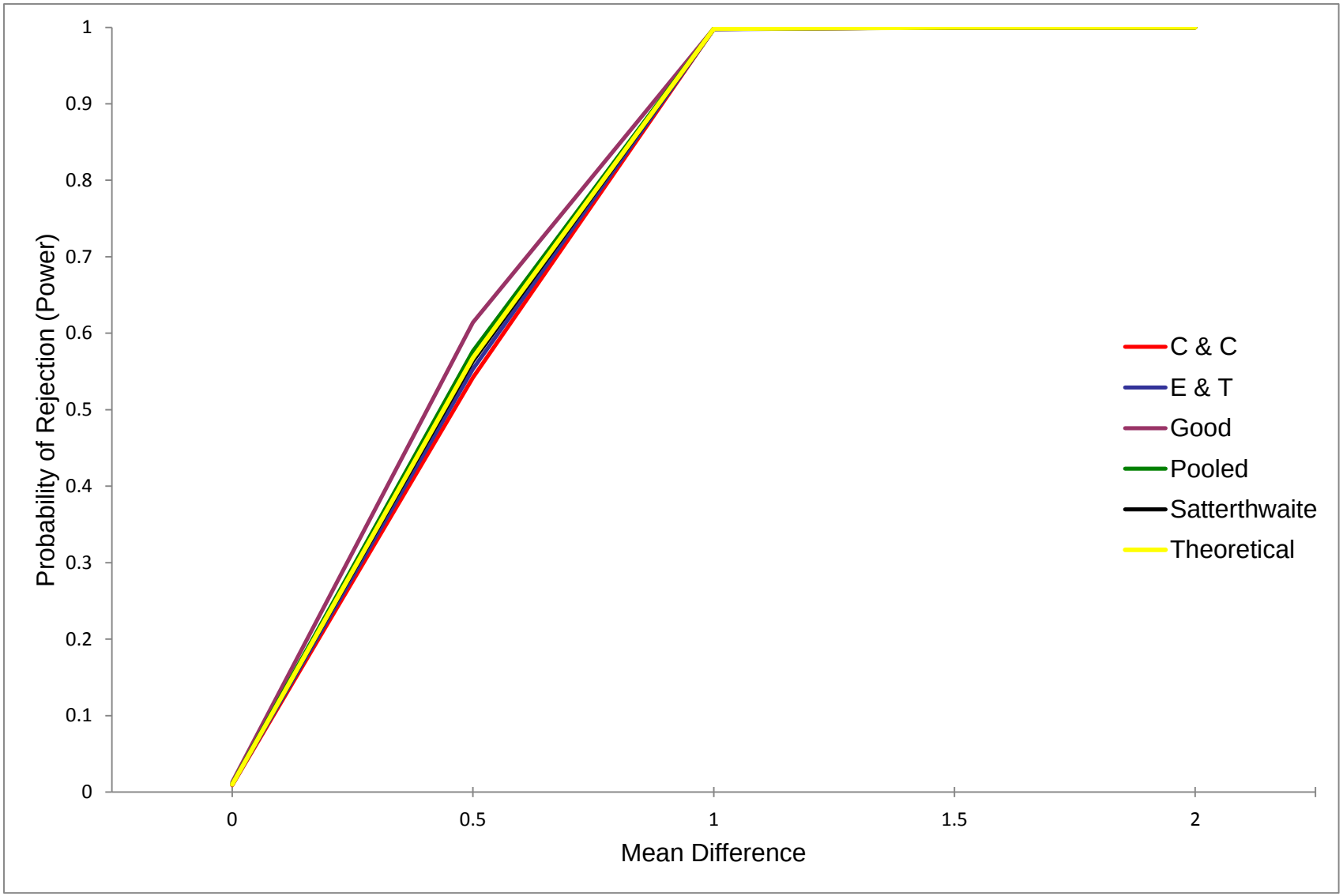

Figure C4. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

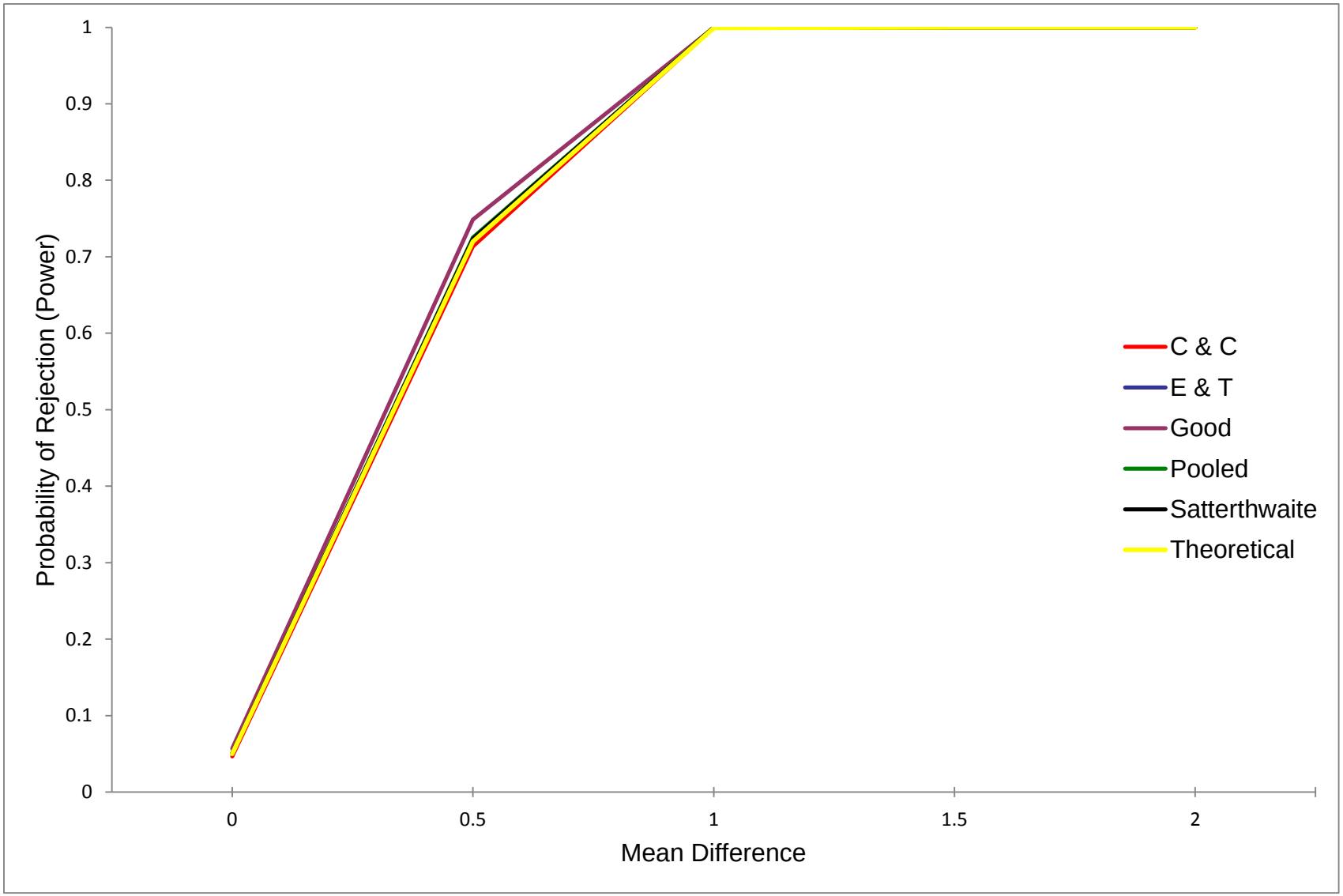

Figure C5. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

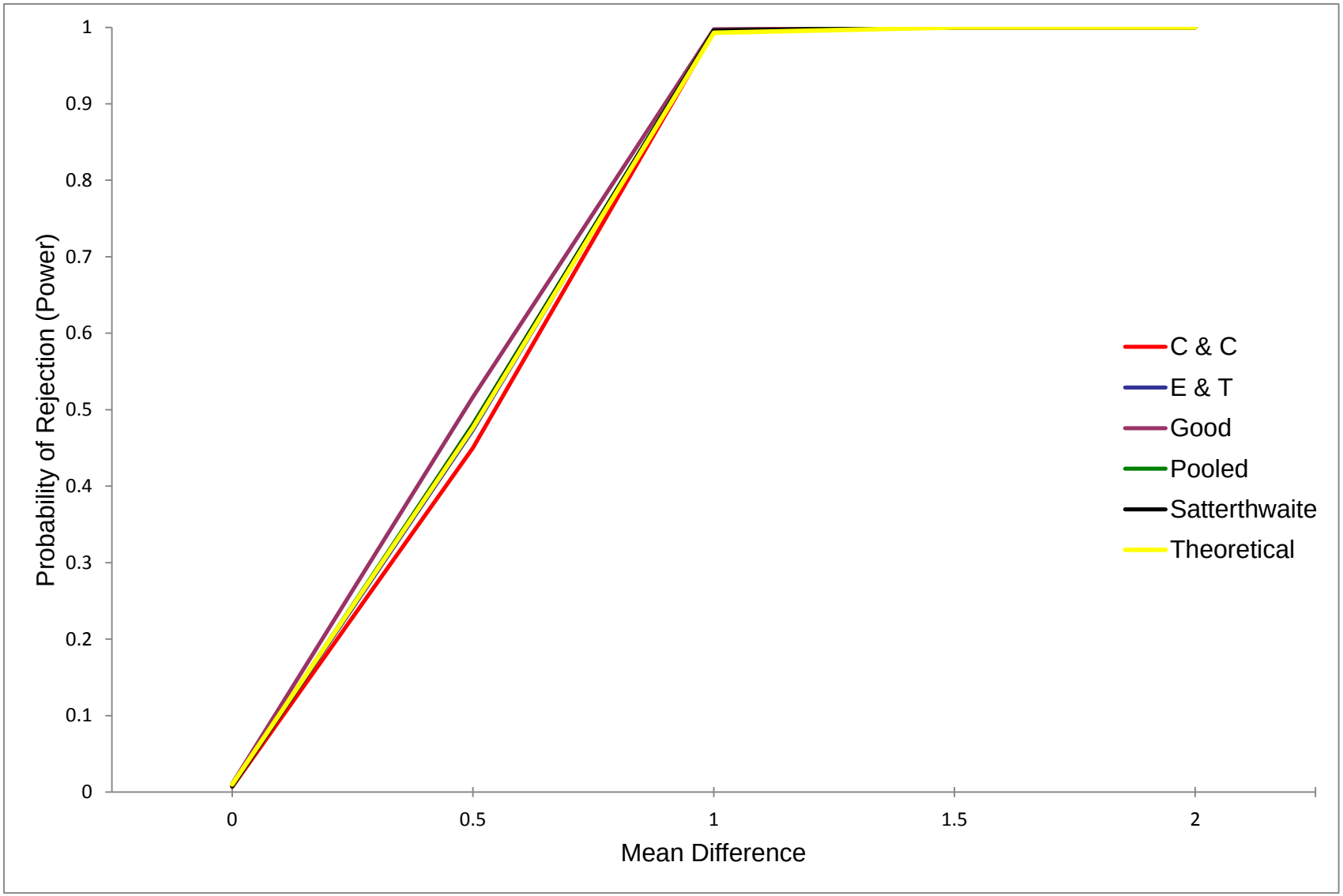

Figure C6. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

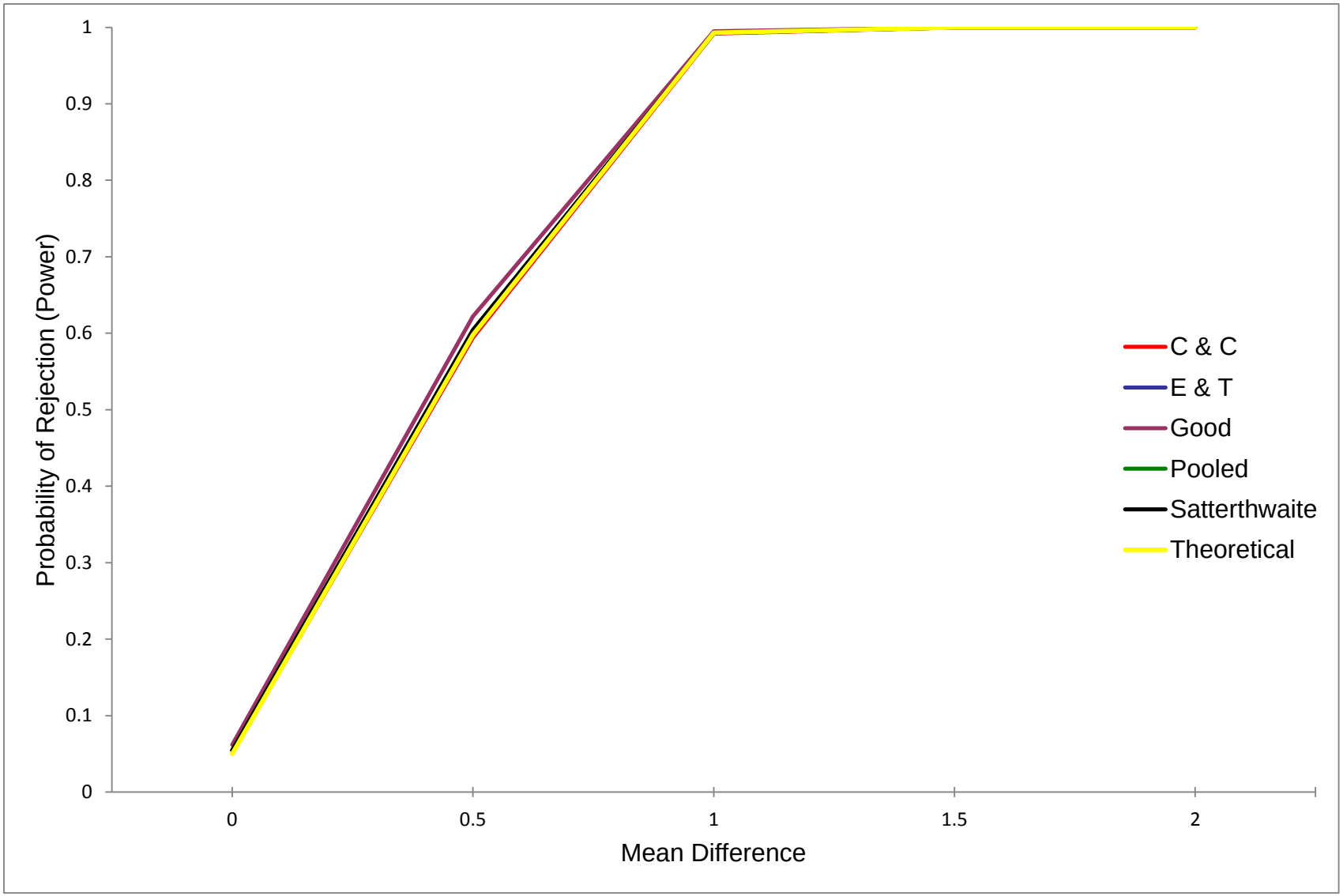

Figure C7. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

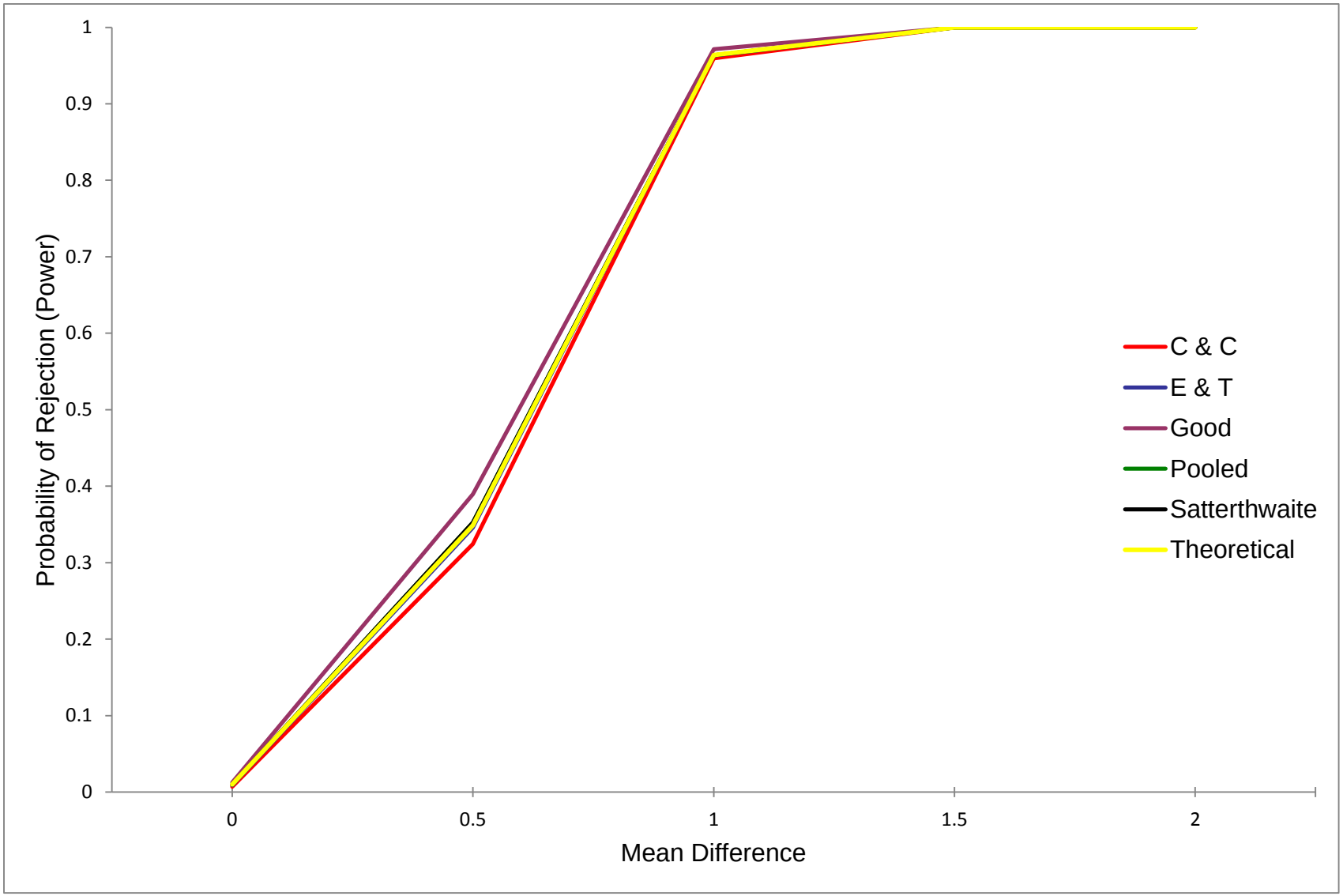

Figure C8. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

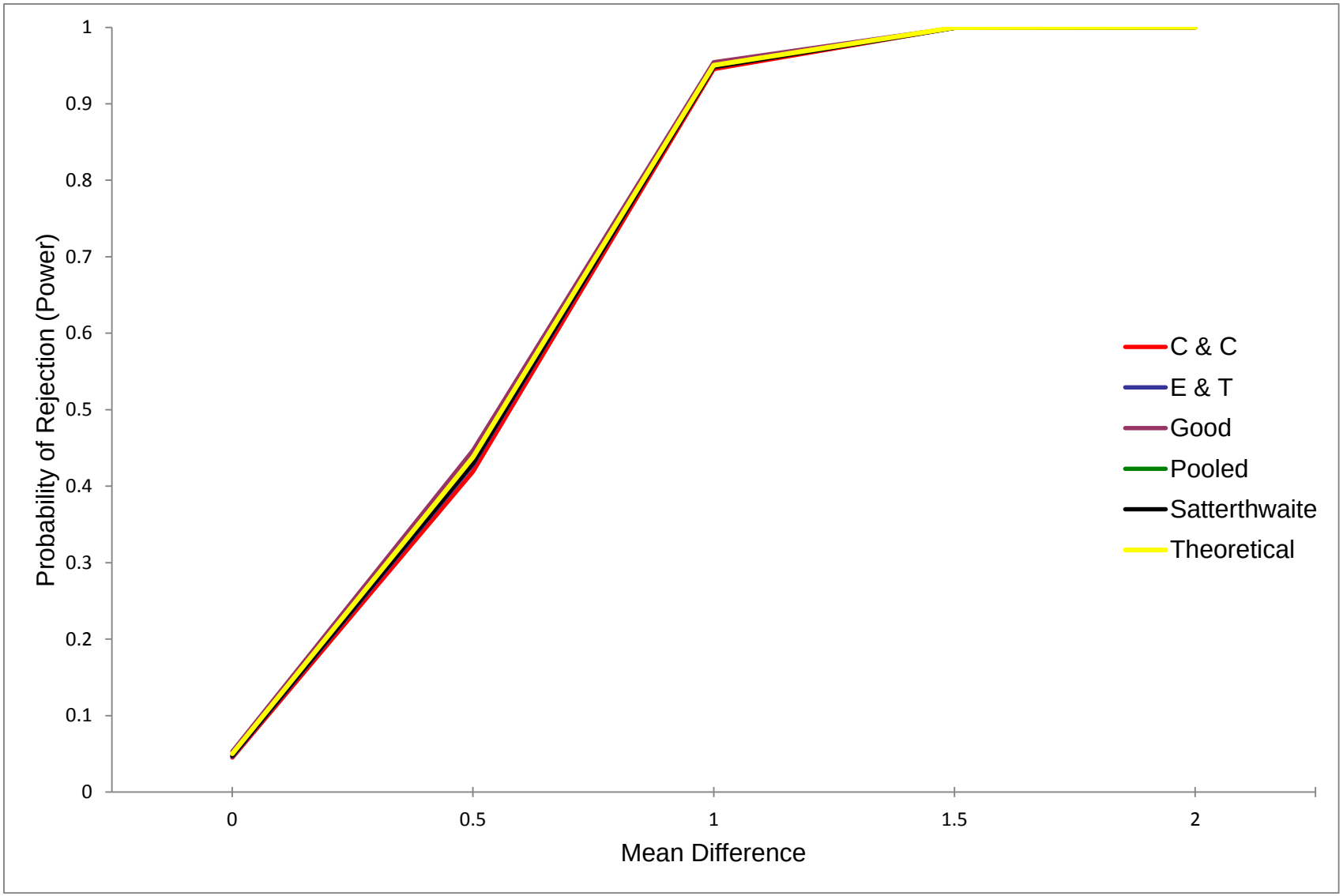

Figure C9. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

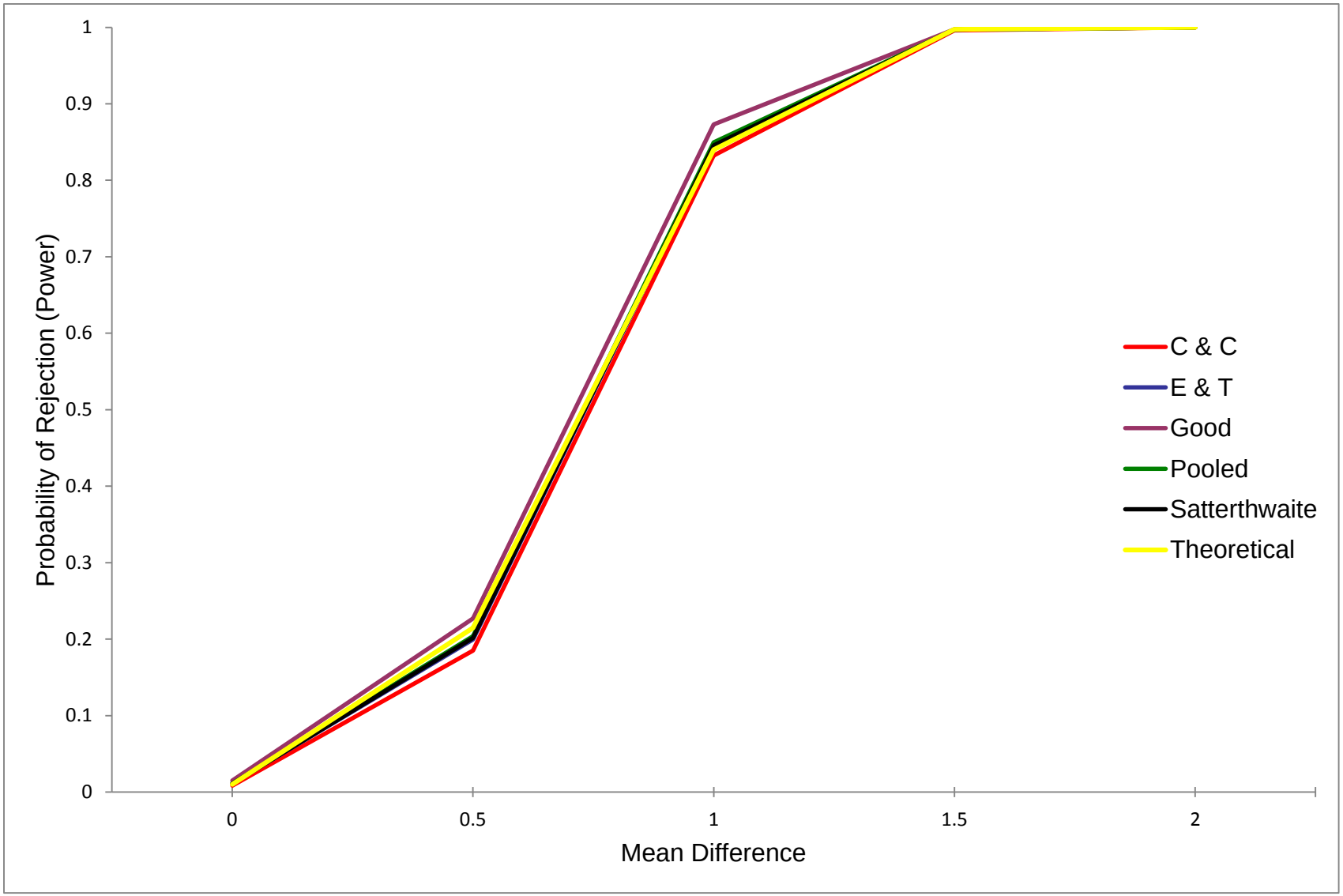

Figure C10 Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

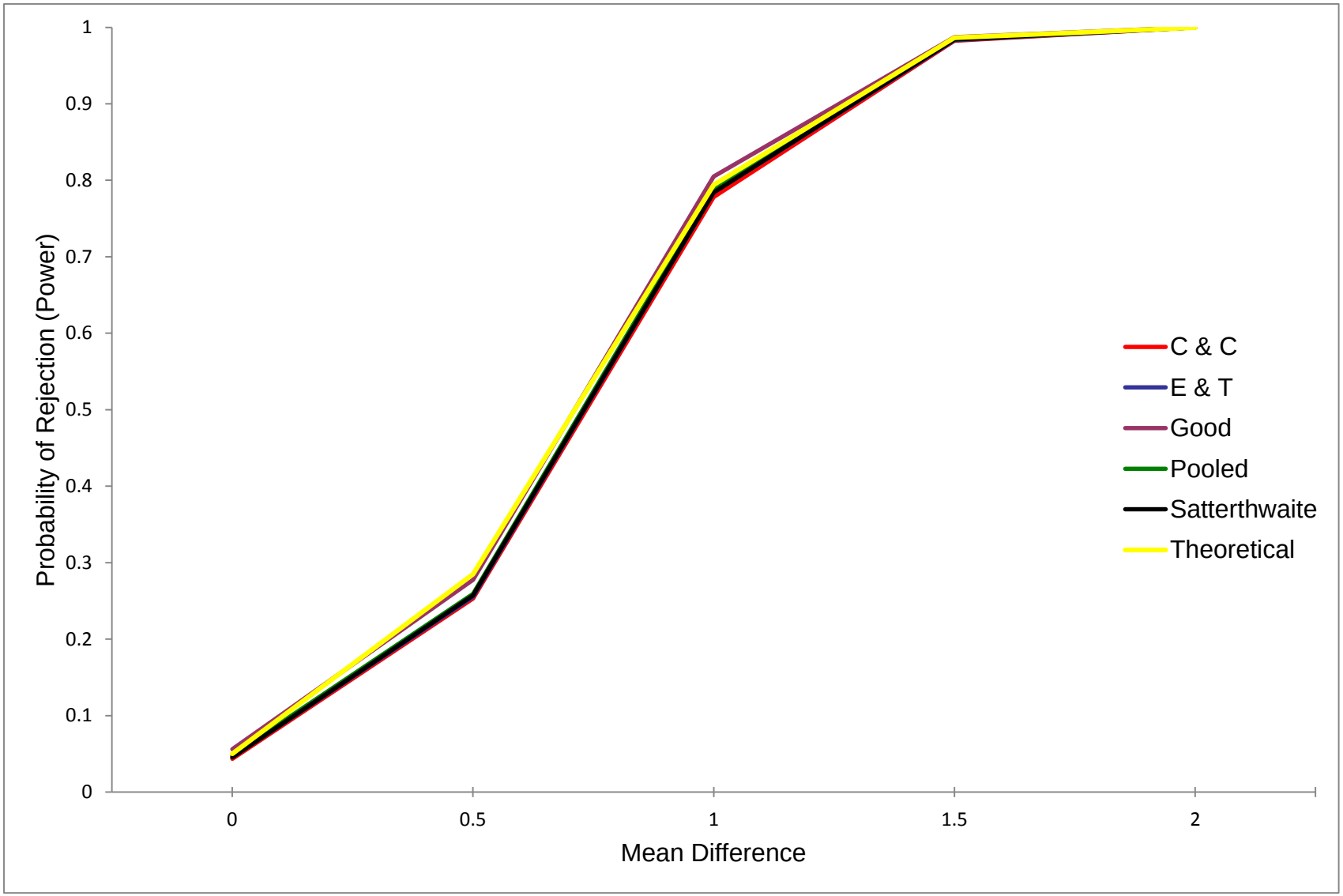

Figure C11. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

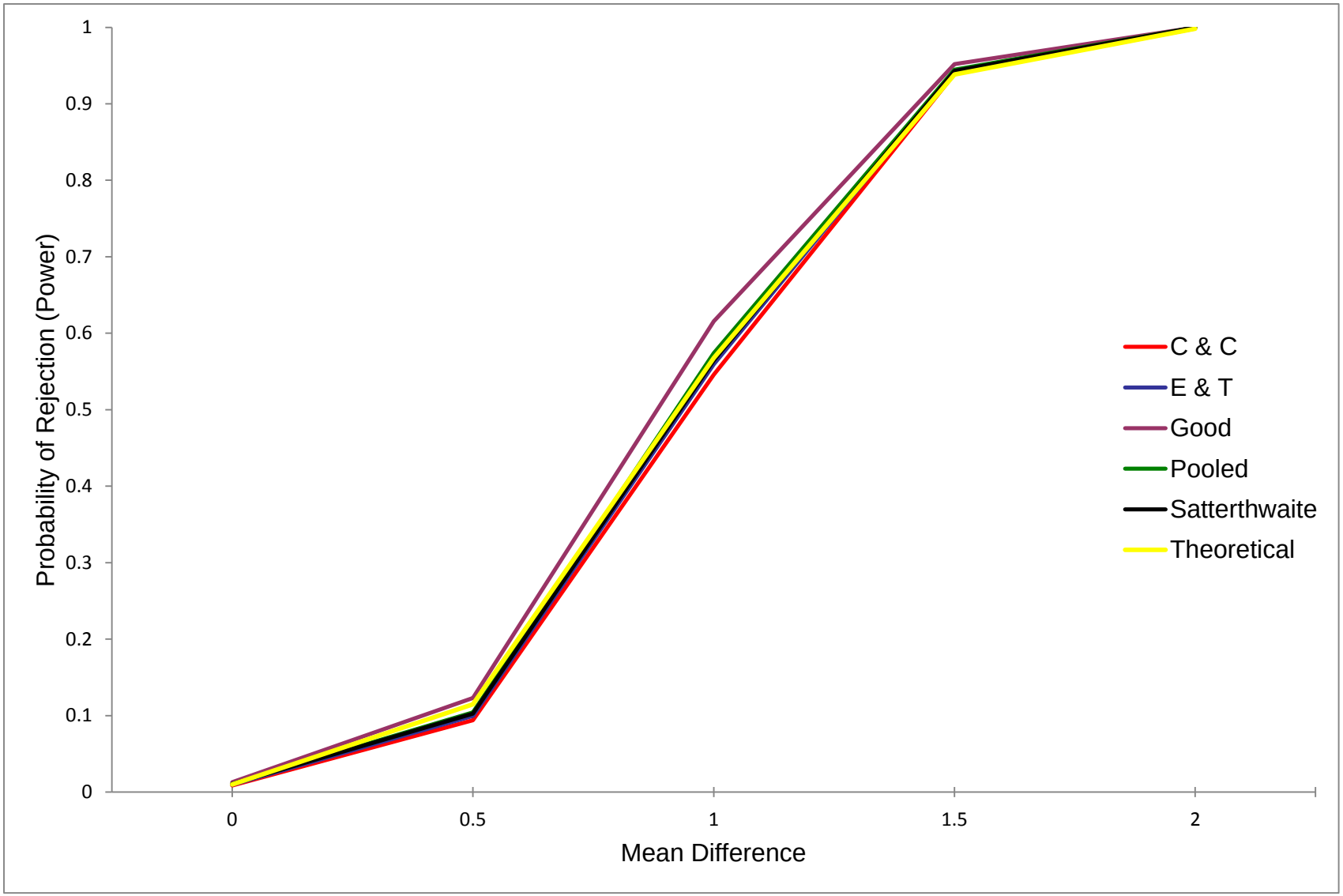

Figure C12. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

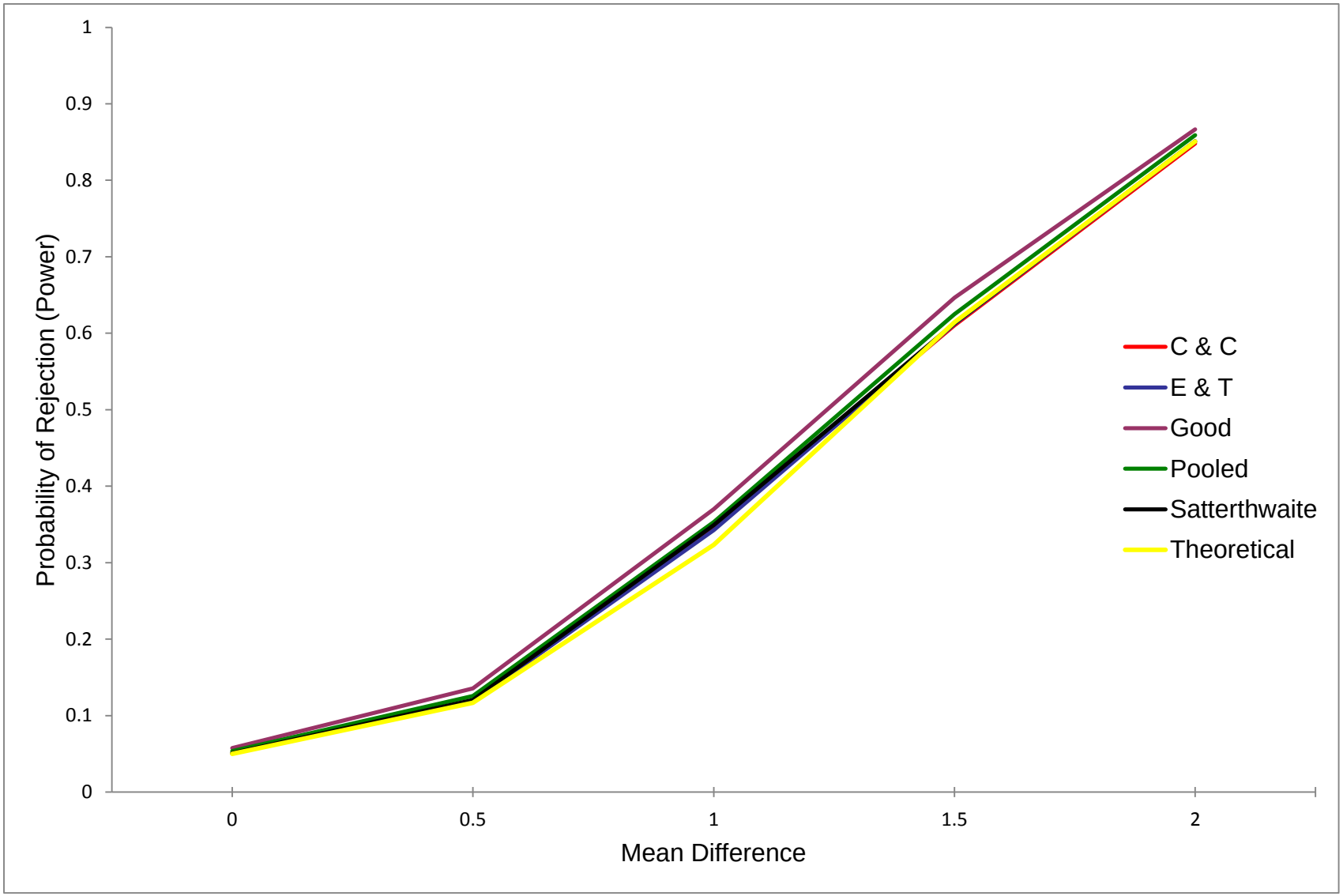

Figure C13. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

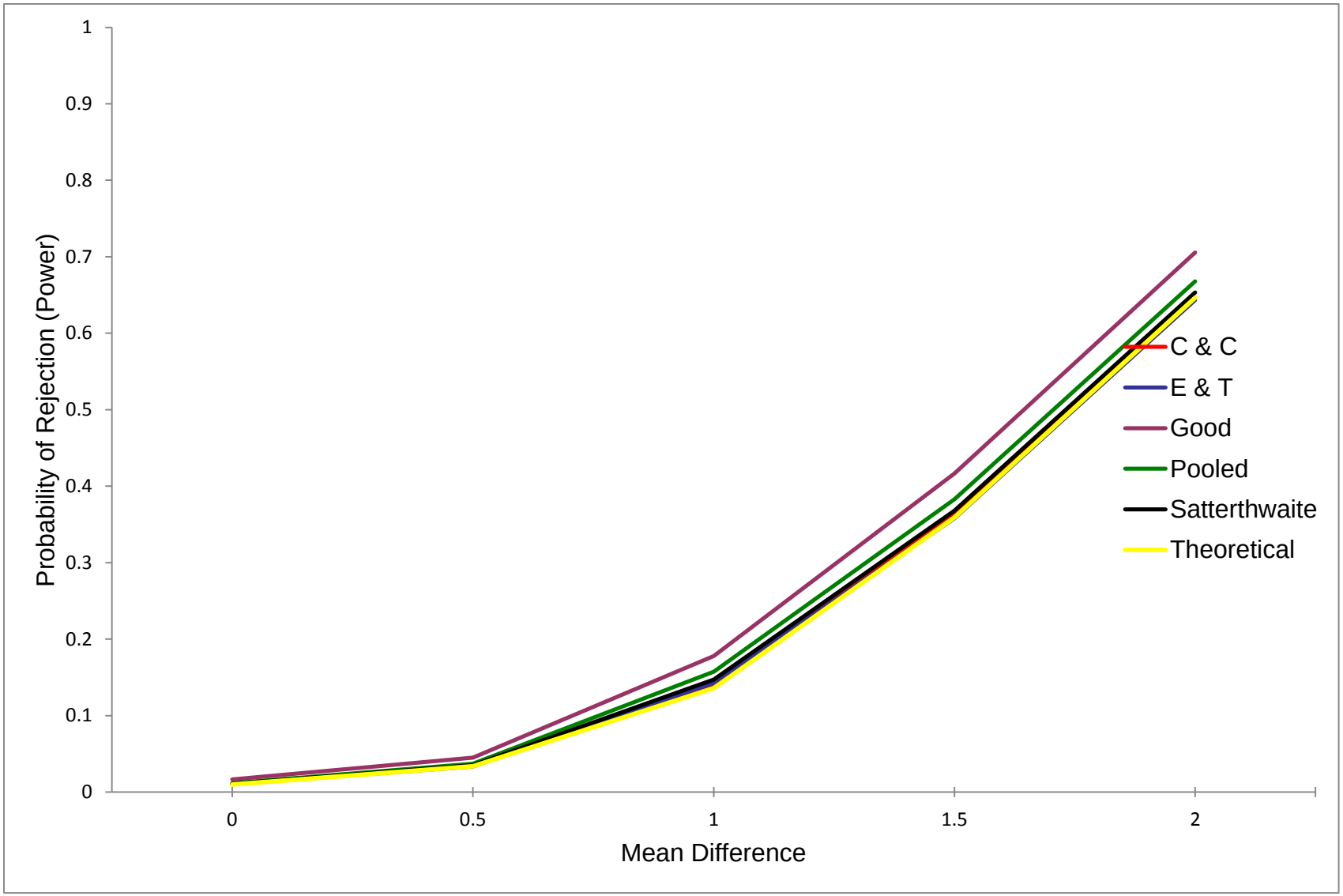

Figure C14. Power curves of all methods for equal group sample sizes when $n_1 = n_2 = 40$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 1.5 (i.e., $n_1 = 40$, $n_2 = 60$)

Table C10

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/16, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0500	0.0105
E & T	0.0530	0.0110
Good	0.0570	0.0150
Pooled	0.0220*	0.0040
Satterthwaite	0.0510	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C11

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/4, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0540	0.0080
E & T	0.0555	0.0105
Good	0.0595	0.0130
Pooled	0.0315*	0.0055
Satterthwaite	0.0560	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C12

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0455	0.0100
E & T	0.0480	0.0120
Good	0.0535	0.0135
Pooled	0.0325*	0.0065
Satterthwaite	0.0470	0.0115

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C13

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0470	0.0115
E & T	0.0520	0.0125
Good	0.0585	0.0150
Pooled	0.0500	0.0125
Satterthwaite	0.0520	0.0130

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C14

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0490	0.0120
E & T	0.0495	0.0120
Good	0.0555	0.0140
Pooled	0.0650	0.0165
Satterthwaite	0.0505	0.0120

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C15

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0470	0.0095
E & T	0.0465	0.0100
Good	0.0560	0.0150
Pooled	0.0740*	0.0235*
Satterthwaite	0.0480	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C16

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0595	0.0090
E & T	0.0580	0.0095
Good	0.0710*	0.0155
Pooled	0.1100*	0.0420*
Satterthwaite	0.0605	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C17

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0500	0.0540	0.0455	0.0470	0.0490	0.0470	0.0595
	0.5	0.9555	0.9105	0.8095	0.6750	0.4655	0.2945	0.1195
	1.0	1.0000	1.0000	1.0000	1.0000	0.9690	0.8070	0.3195
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9860	0.6205
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8645
Efron & Tibshirani	0.0	0.0530	0.0555	0.0480	0.0520	0.0495	0.0465	0.0580
	0.5	0.9560	0.9135	0.8145	0.6860	0.4710	0.2985	0.1175
	1.0	1.0000	1.0000	1.0000	1.0000	0.9700	0.8110	0.3180
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9860	0.6190
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8620
Good	0.0	0.0570	0.0595	0.0535	0.0585	0.0555	0.0560	0.0710
	0.5	0.9610	0.9220	0.8260	0.7065	0.4965	0.3270	0.1405
	1.0	1.0000	1.0000	1.0000	1.0000	0.9745	0.8335	0.3520
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9865	0.6580
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8815
Pooled	0.0	0.0220	0.0315	0.0325	0.0500	0.0650	0.0740	0.1100
	0.5	0.9100	0.8650	0.7825	0.6915	0.5255	0.3890	0.2115
	1.0	1.0000	1.0000	1.0000	1.0000	0.9780	0.8695	0.4610
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9920	0.7510
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9280
Satterthwaite	0.0	0.0510	0.0560	0.0470	0.0520	0.0505	0.0480	0.0605
	0.5	0.9555	0.9140	0.8155	0.6855	0.4725	0.3000	0.1195
	1.0	1.0000	1.0000	1.0000	1.0000	0.9695	0.8110	0.3215
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9860	0.6225
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8655

Table C18

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 60$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0105	0.0080	0.0100	0.0115	0.0120	0.0095	0.0090
	0.5	0.8455	0.7520	0.6030	0.4145	0.2170	0.1130	0.0315
	1.0	1.0000	1.0000	0.9995	0.9870	0.8835	0.5795	0.1390
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9535	0.3535
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.6680
Efron & Tibshirani	0.0	0.0110	0.0105	0.0120	0.0125	0.0120	0.0100	0.0095
	0.5	0.8450	0.7585	0.6235	0.4385	0.2260	0.1140	0.0290
	1.0	1.0000	1.0000	0.9995	0.9885	0.8890	0.5865	0.1375
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9520	0.3430
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.6620
Good	0.0	0.0150	0.0130	0.0135	0.0150	0.0140	0.0150	0.0155
	0.5	0.8700	0.7870	0.6540	0.4750	0.2600	0.1435	0.0435
	1.0	1.0000	1.0000	1.0000	0.9930	0.9140	0.6400	0.1800
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9650	0.4170
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.7220
Pooled	0.0	0.0040	0.0055	0.0065	0.0125	0.0165	0.0235	0.0420
	0.5	0.6985	0.6430	0.5570	0.4400	0.3000	0.1985	0.0905
	1.0	1.0000	1.0000	0.9990	0.9900	0.9285	0.7135	0.2595
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9775	0.5500
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8250
Satterthwaite	0.0	0.0110	0.0105	0.0115	0.0130	0.0120	0.0105	0.0090
	0.5	0.8495	0.7605	0.6255	0.4415	0.2345	0.1210	0.0325
	1.0	1.0000	1.0000	0.9995	0.9890	0.8925	0.5905	0.1405
	1.5	1.0000	1.0000	1.0000	1.0000	0.9990	0.9550	0.3545
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	0.6700

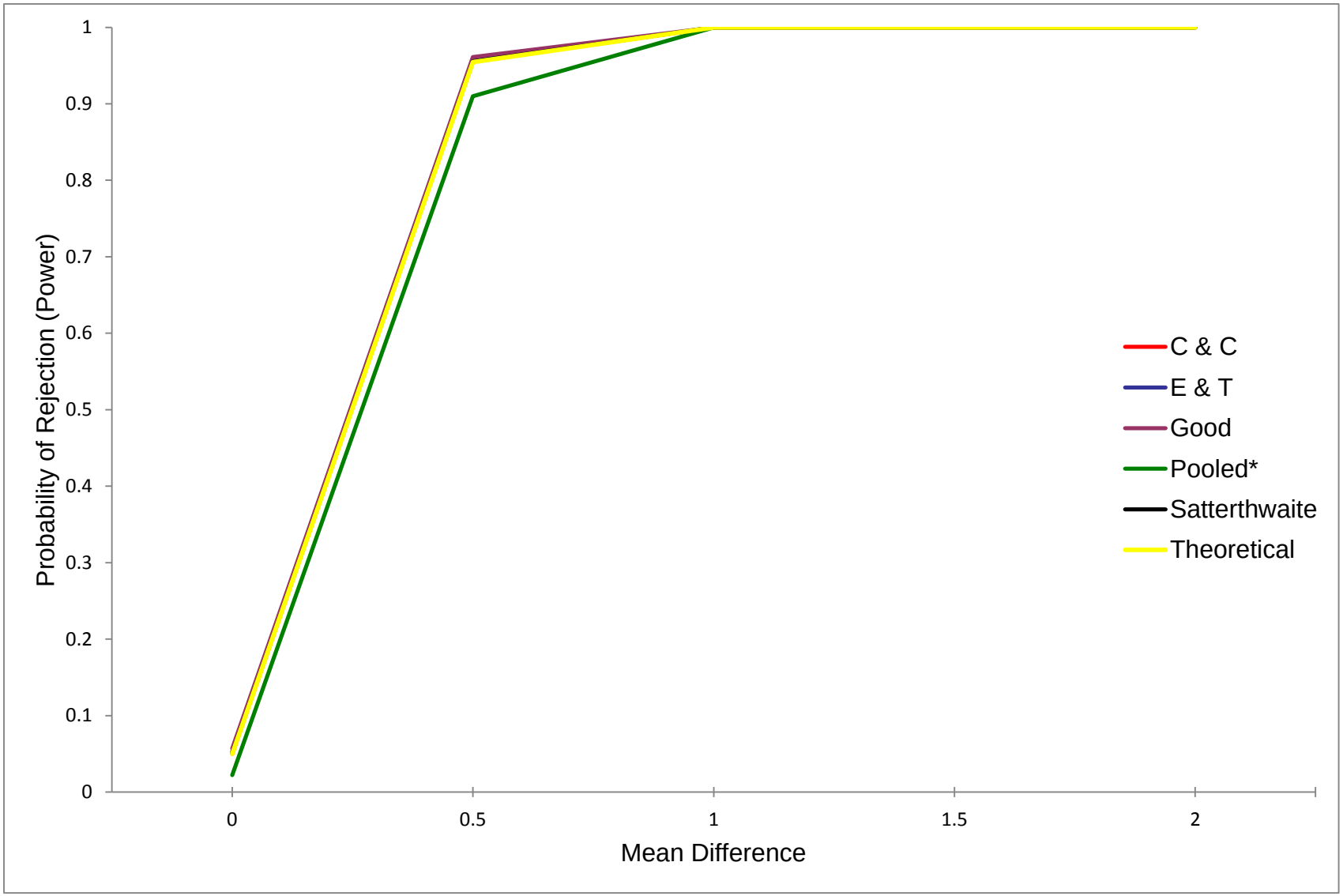

Figure C15. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

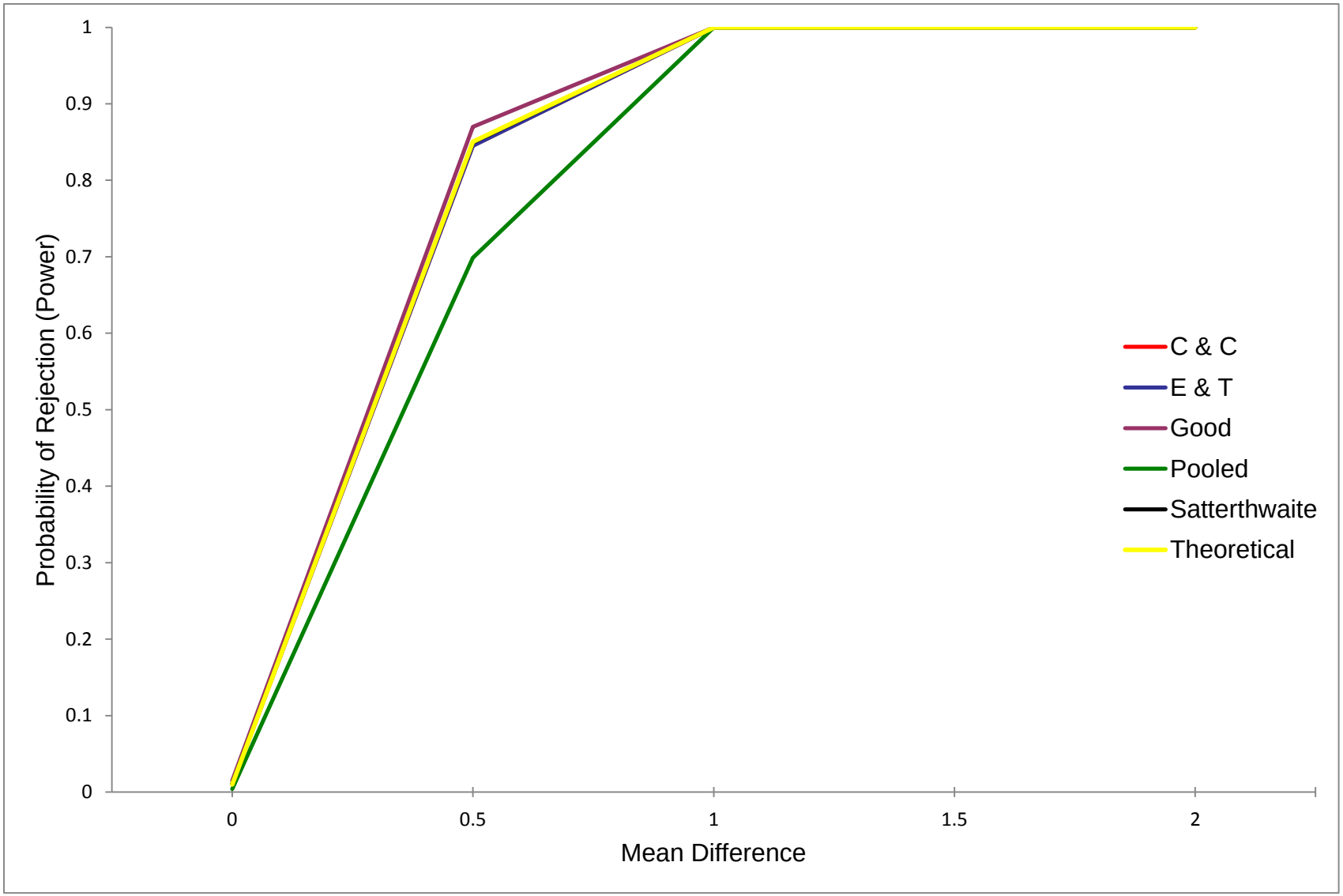

Figure C16. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

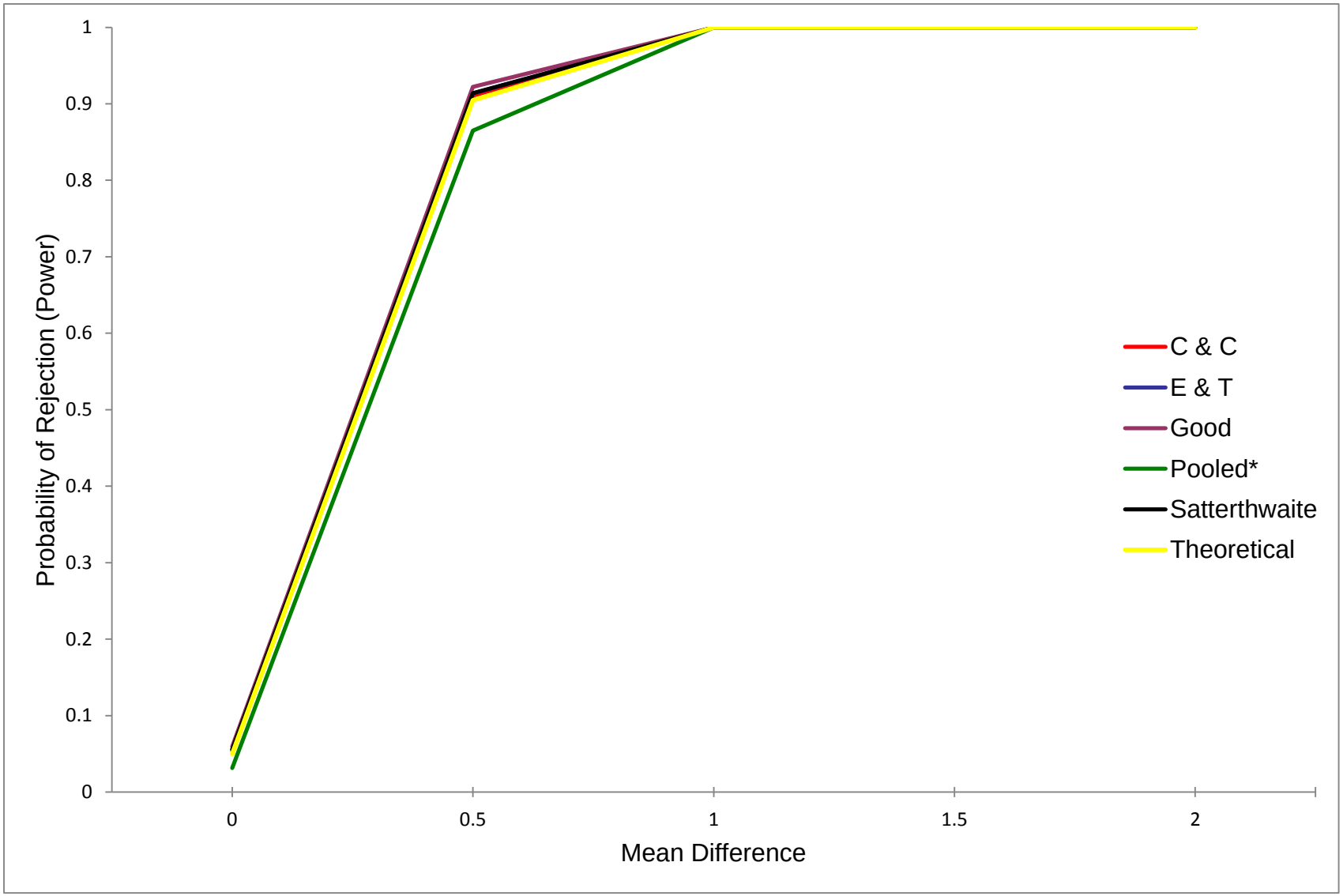

Figure C17. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

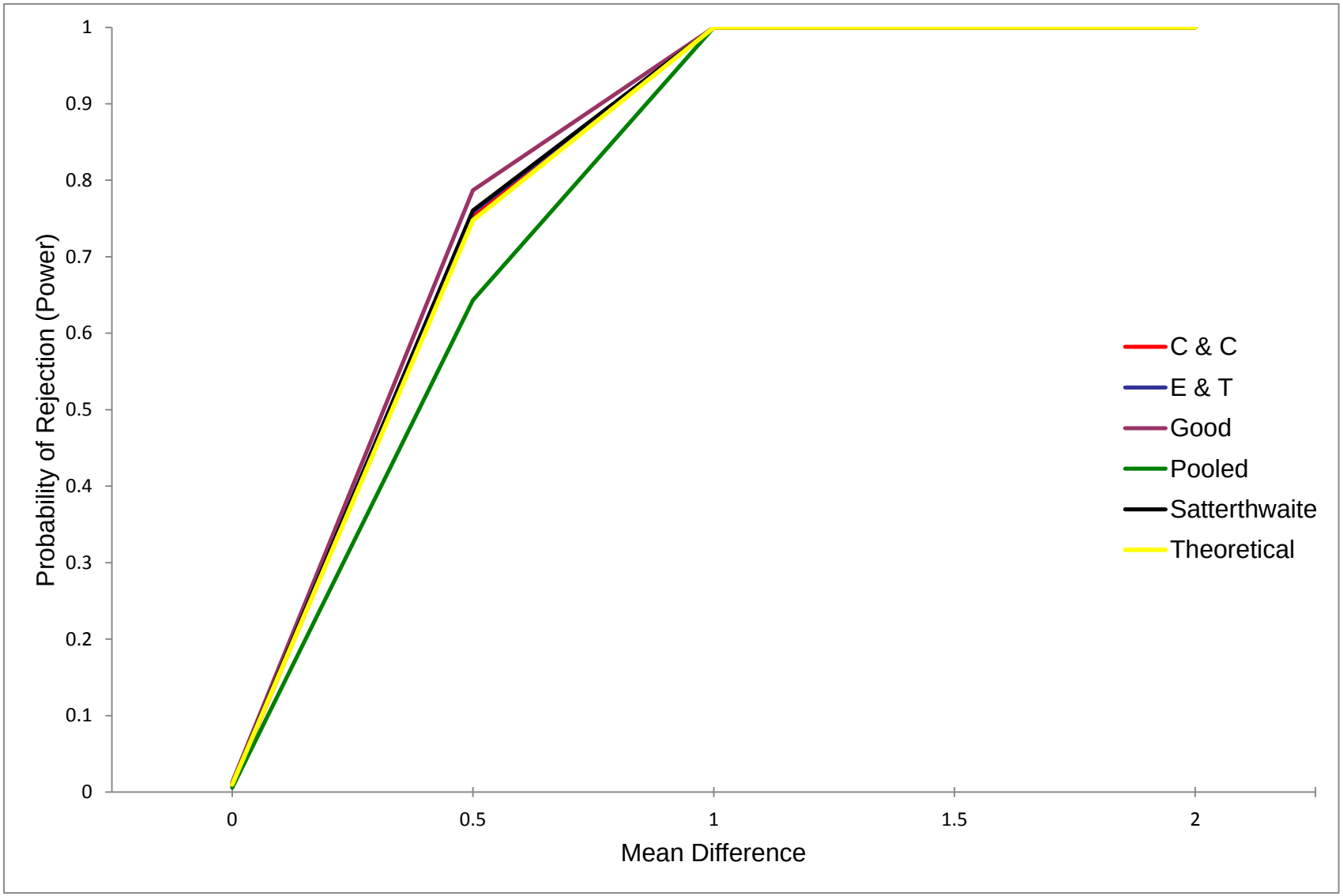

Figure C18. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

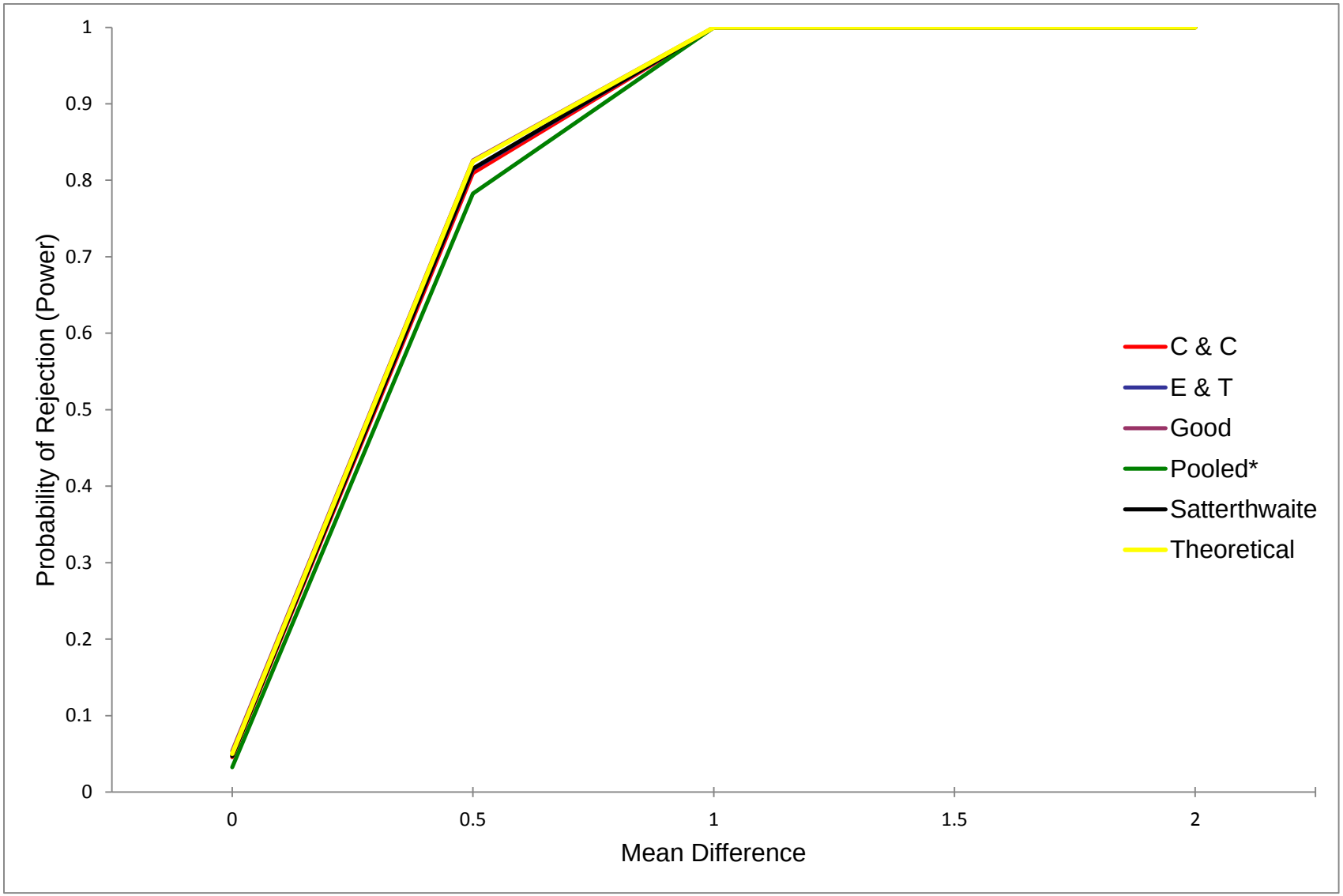

Figure C19. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

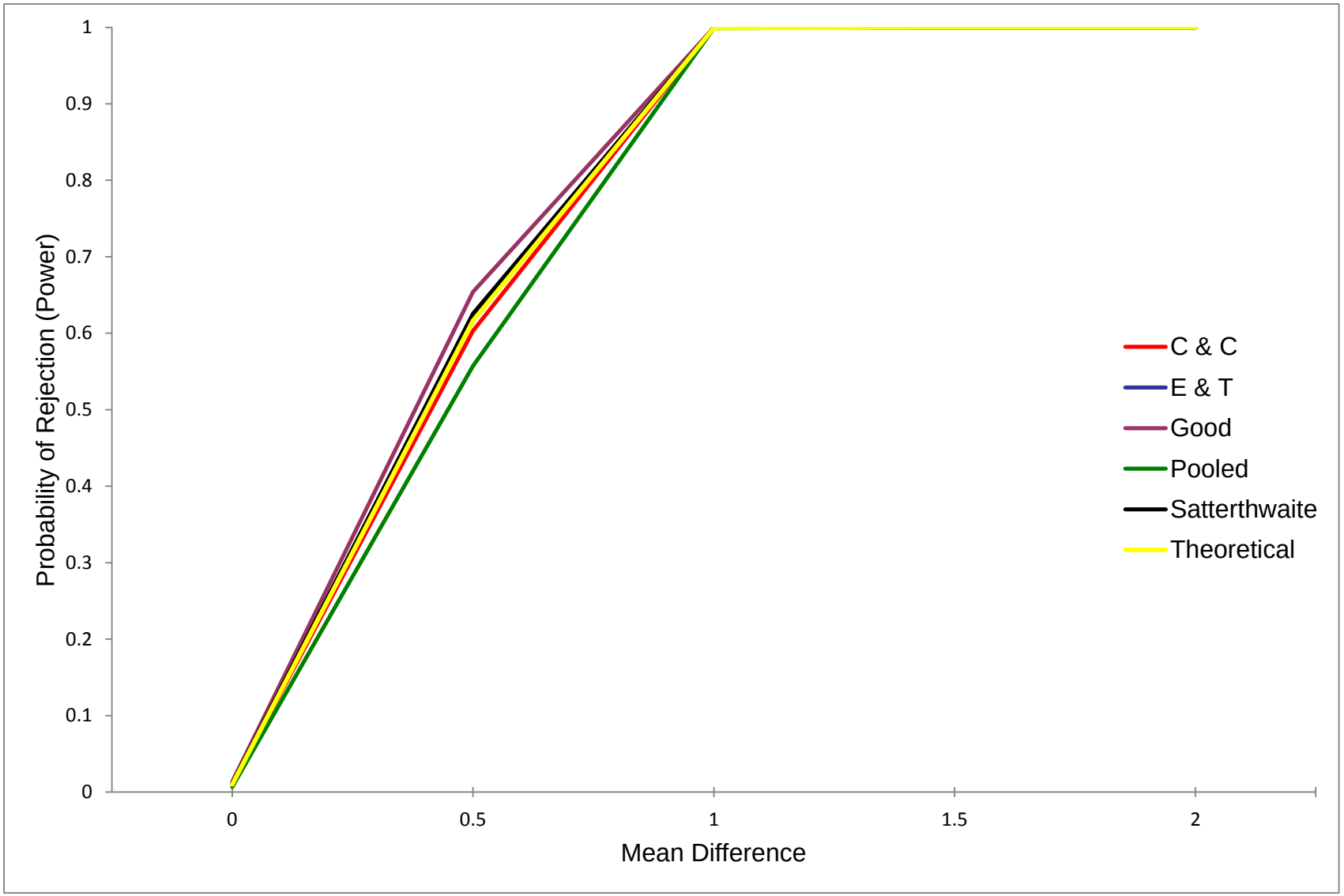

Figure C20. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

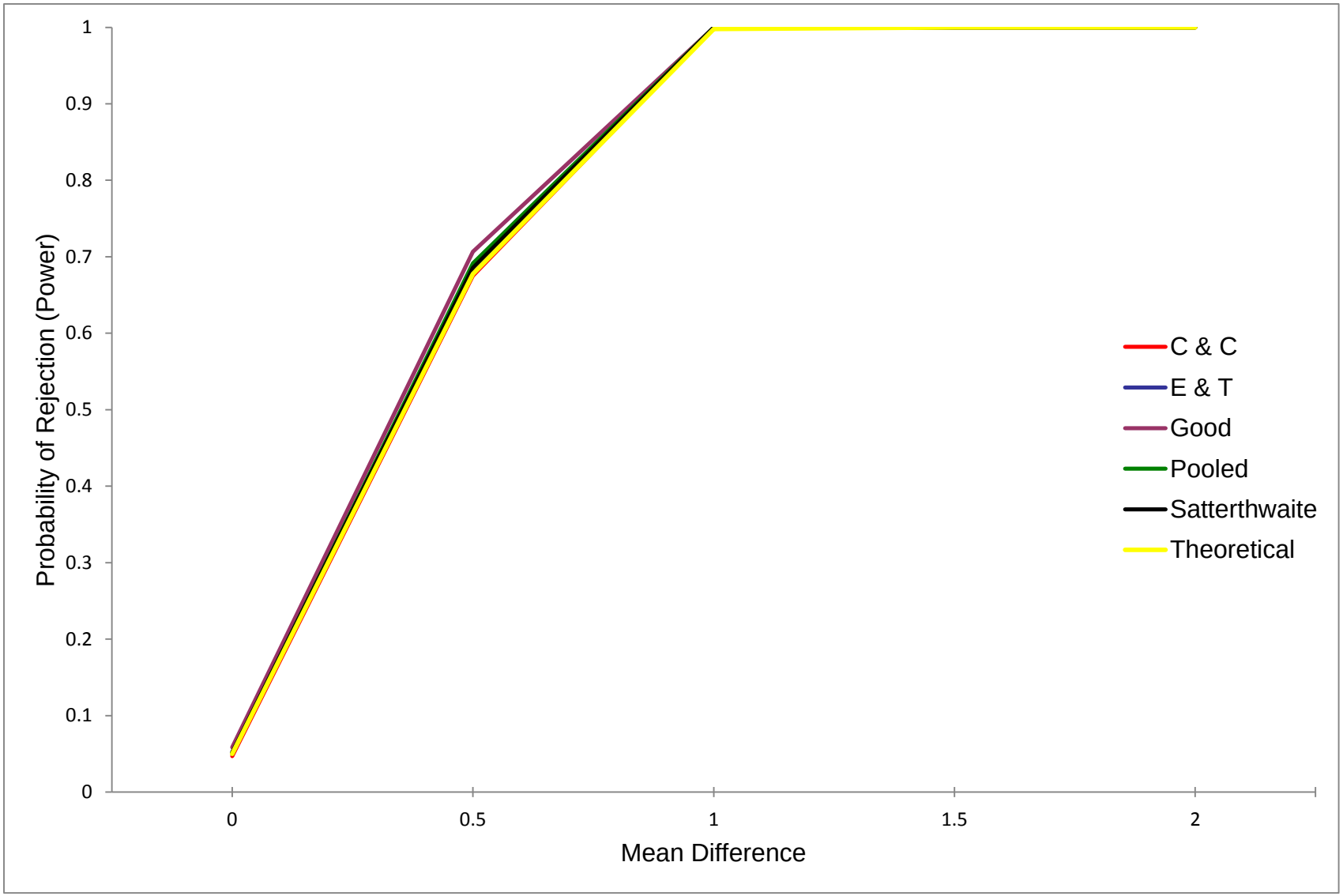

Figure C21. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

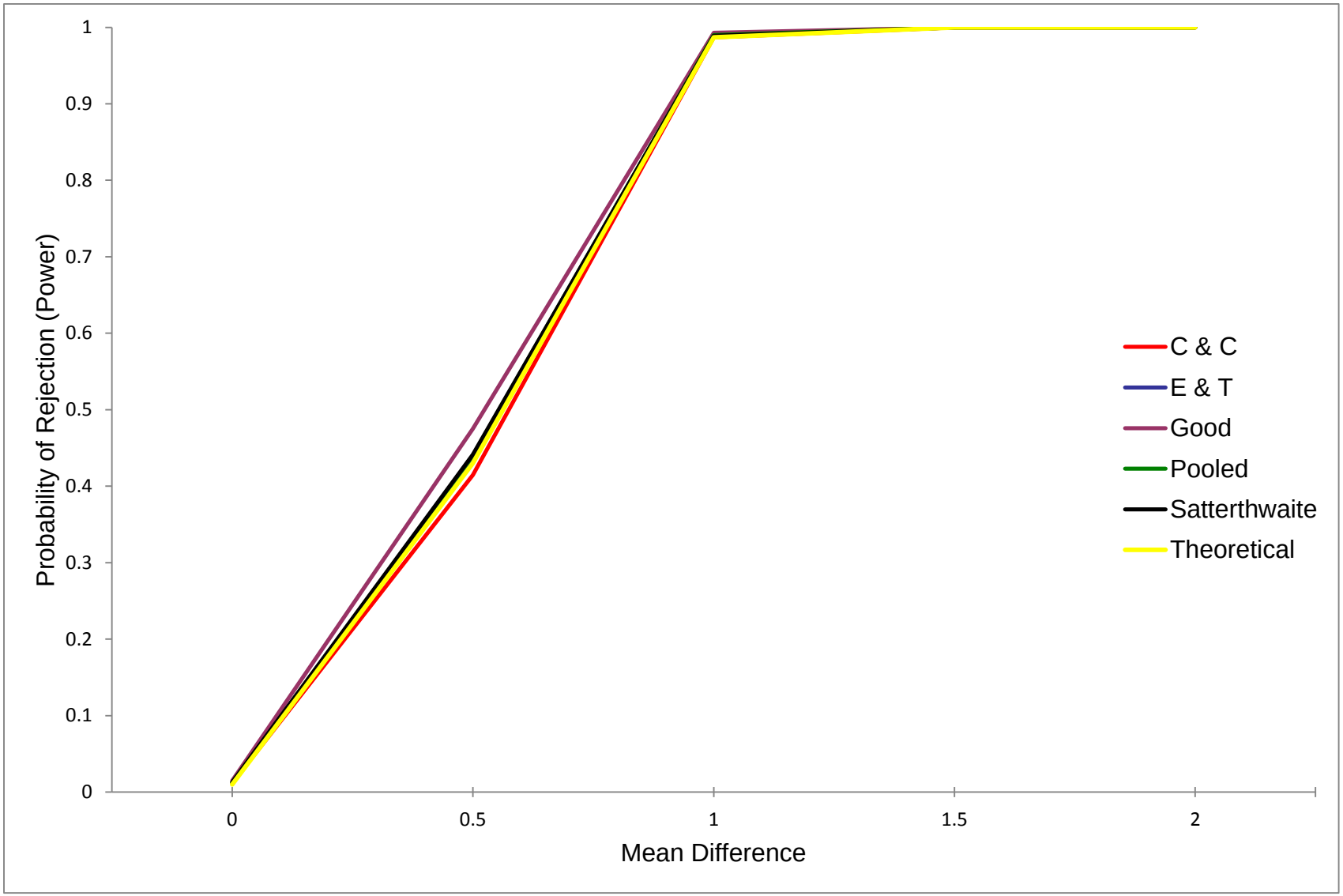

Figure C22. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

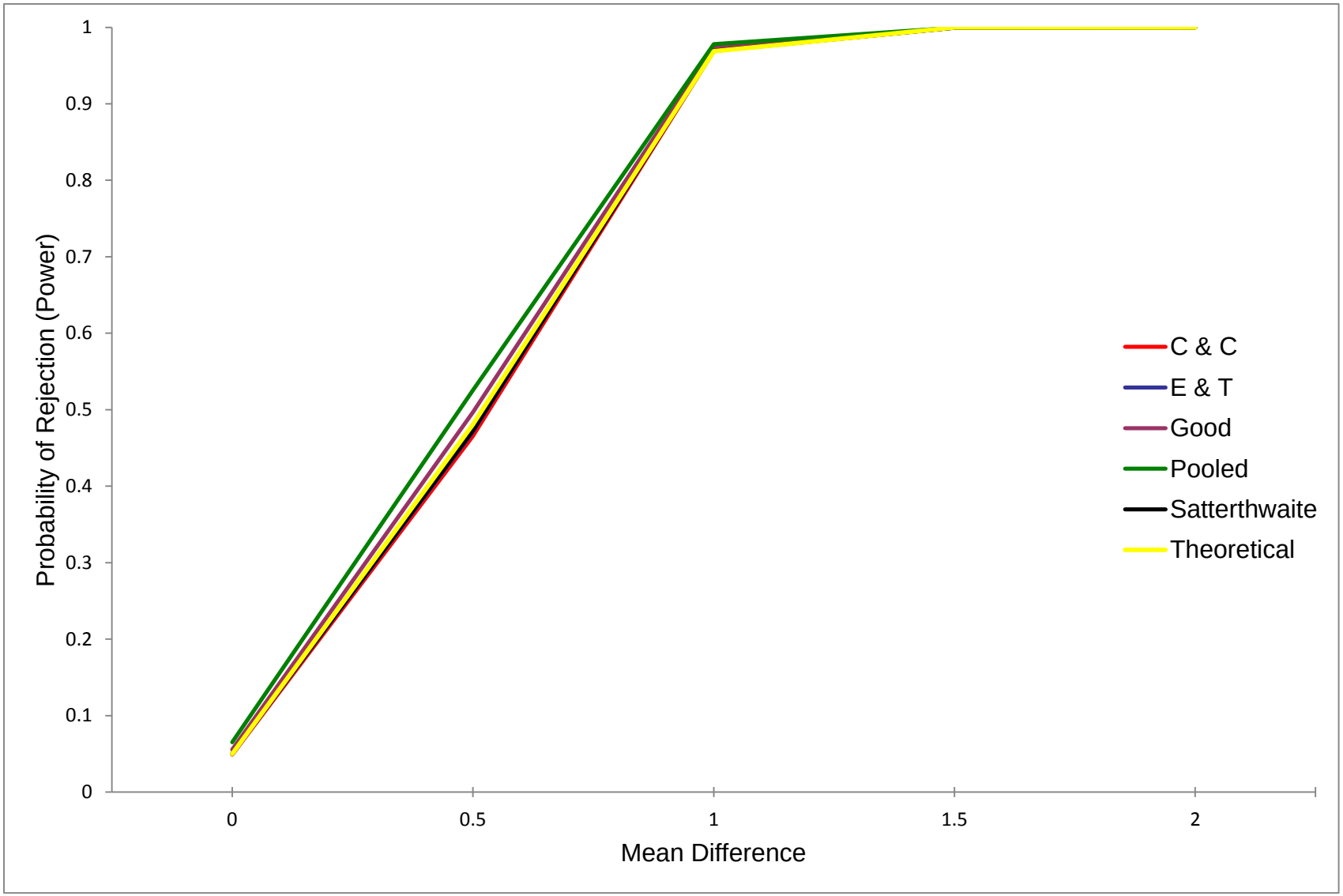

Figure C23. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

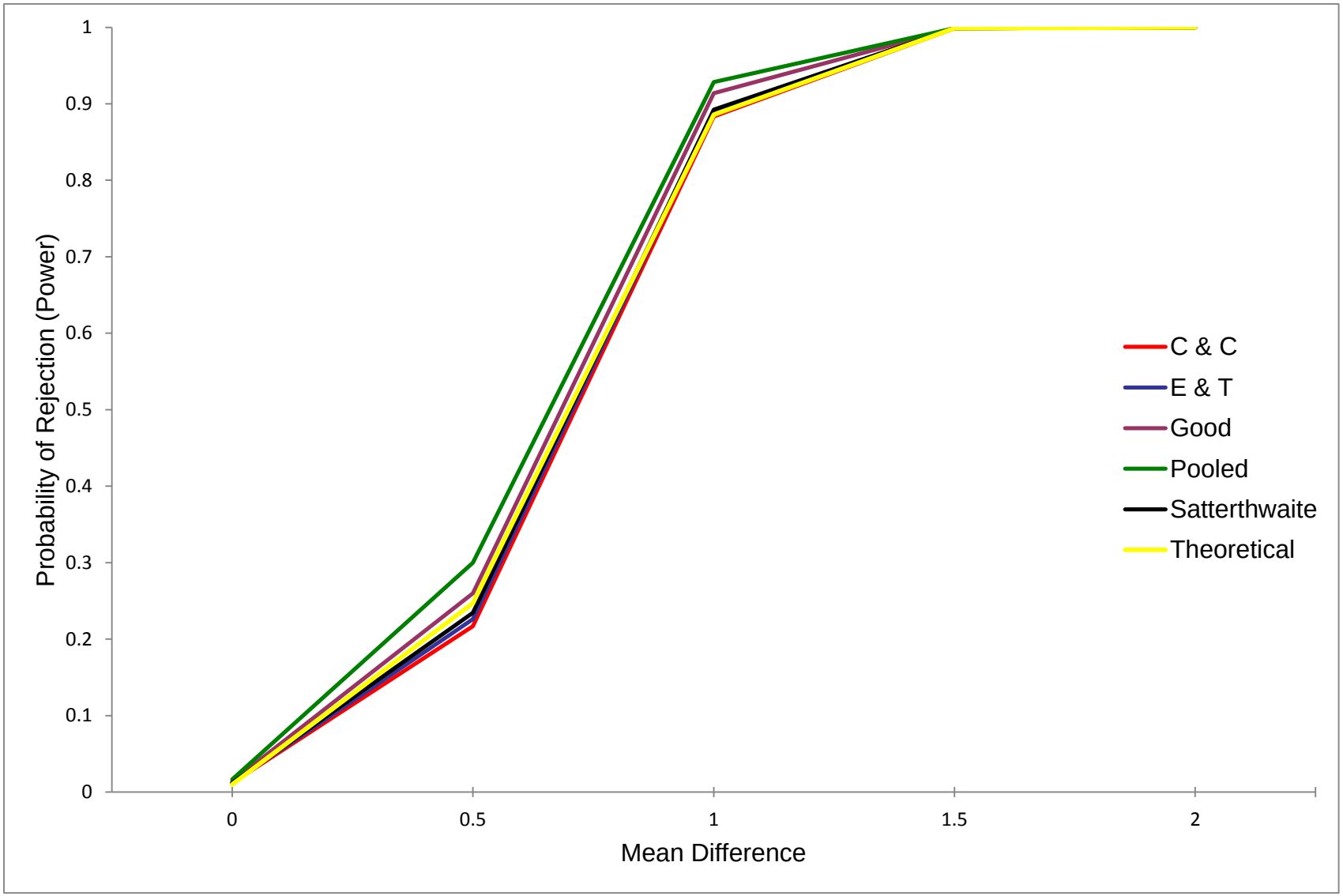

Figure C24. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

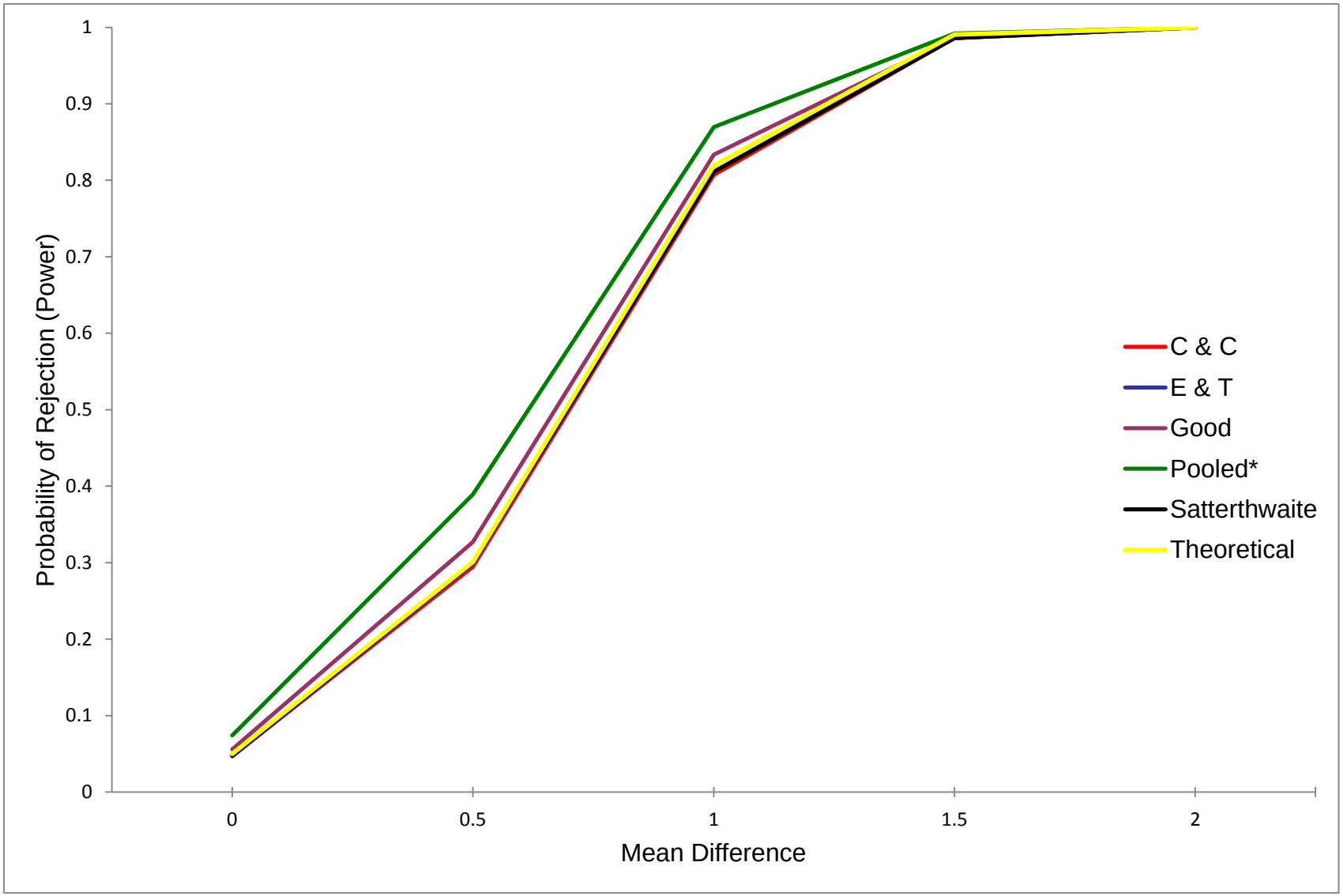

Figure C25. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

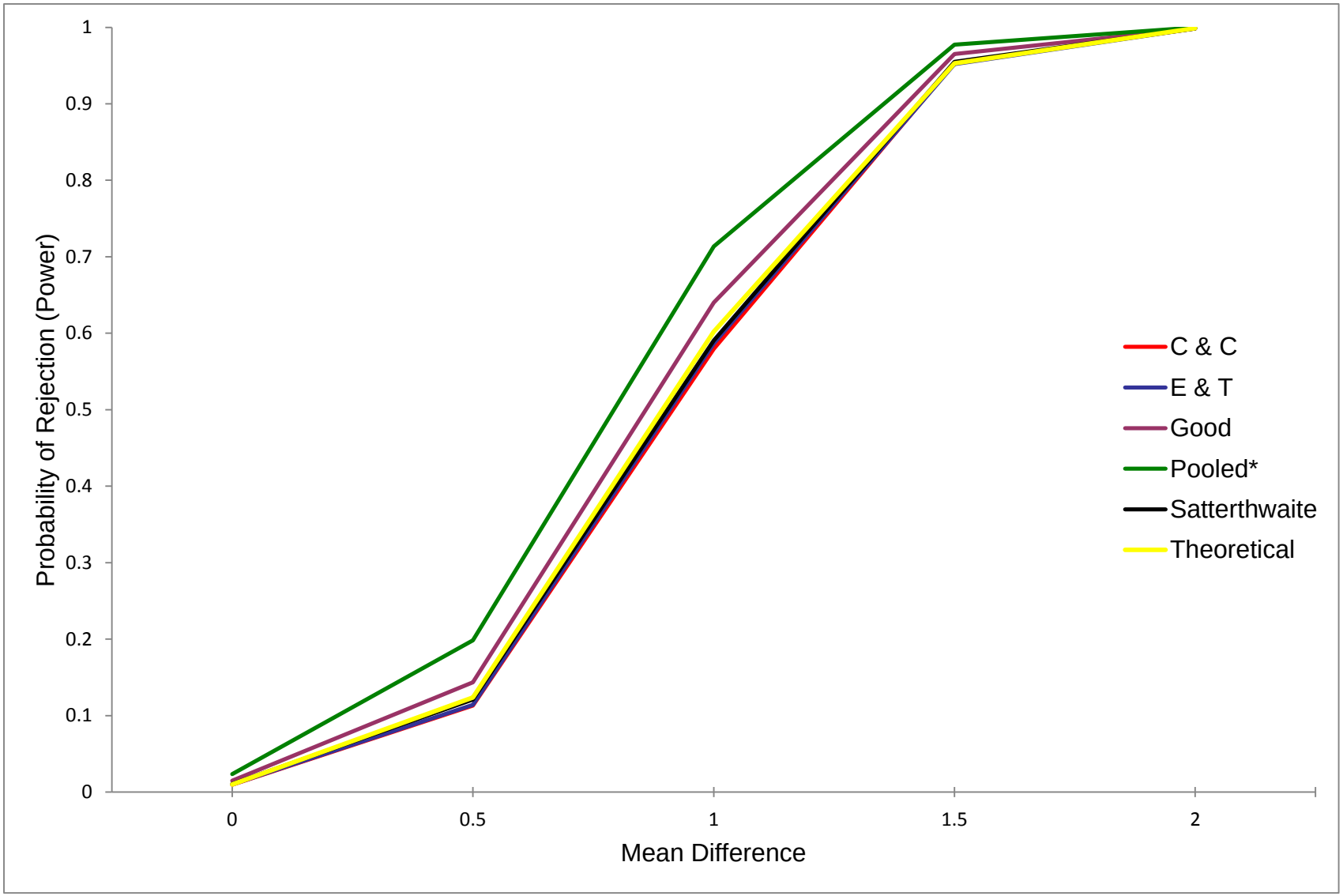

Figure C26. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

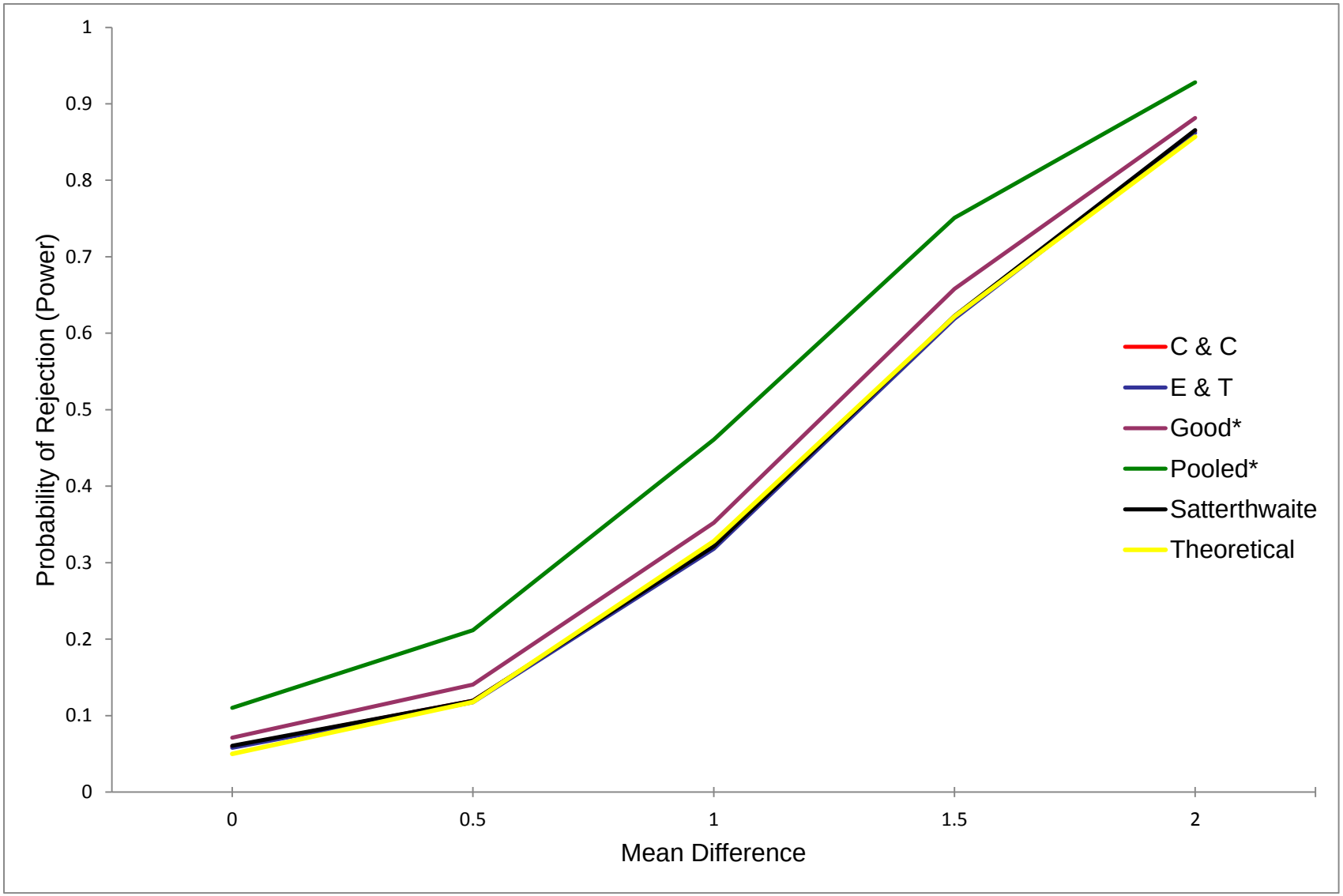

Figure C27. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

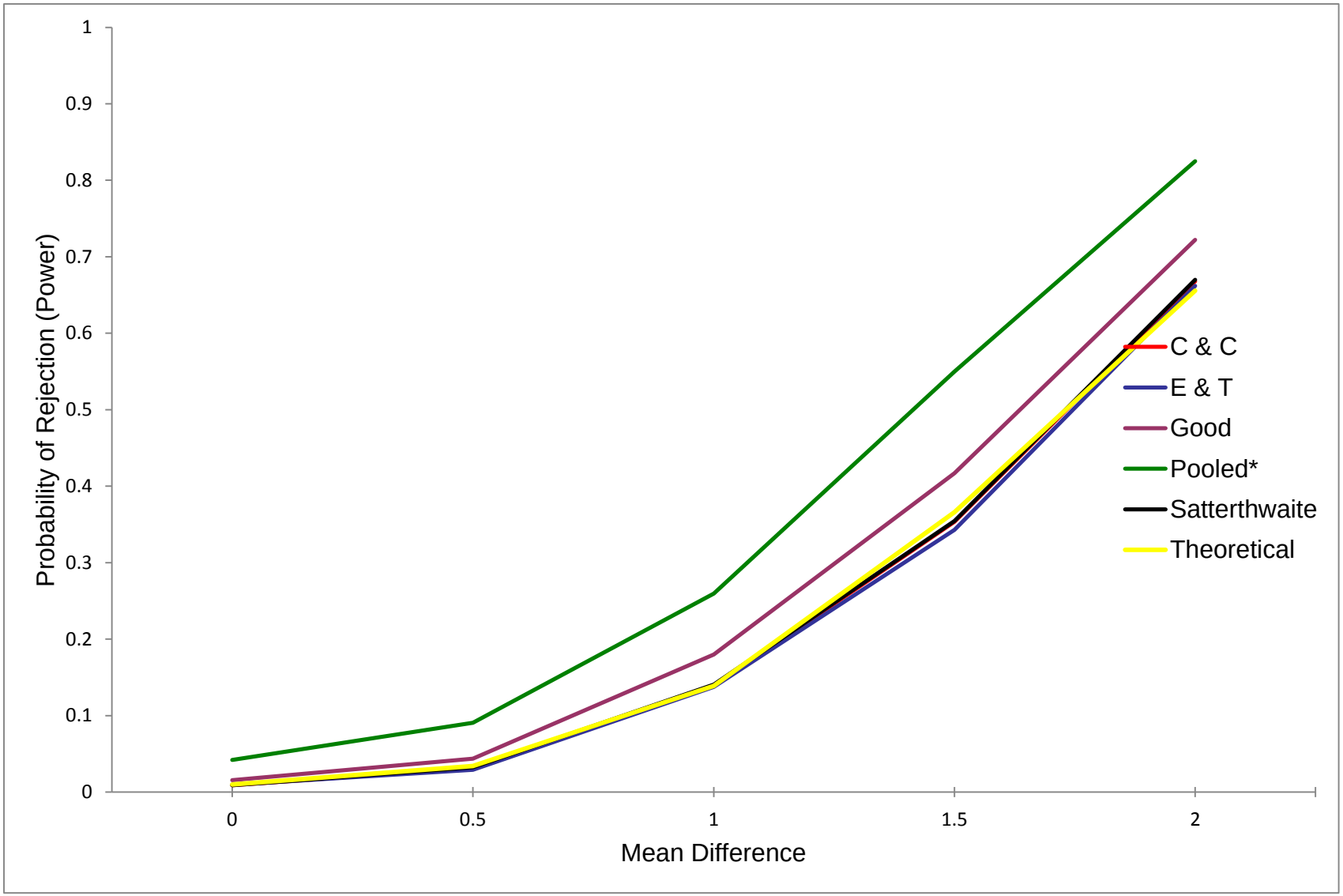

Figure C28. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 60$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 3.0 (i.e., $n_1 = 40$, $n_2 = 120$)

Table C19

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0430	0.0085
E & T	0.0455	0.0090
Good	0.0505	0.0105
Pooled	0.0020*	<.0005*
Satterthwaite	0.0455	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C20

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0455	0.0090
E & T	0.0490	0.0105
Good	0.0535	0.0120
Pooled	0.0060*	<.0005*
Satterthwaite	0.0490	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C21

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0505	0.0095
E & T	0.0535	0.0110
Good	0.0585	0.0125
Pooled	0.0215*	0.0025*
Satterthwaite	0.0530	0.0105

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C22

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0420	0.0080
E & T	0.0455	0.0090
Good	0.0545	0.0120
Pooled	0.0450	0.0090
Satterthwaite	0.0460	0.0090

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C23

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0505	0.0085
E & T	0.0500	0.0095
Good	0.0595	0.0145
Pooled	0.0945*	0.0260*
Satterthwaite	0.0505	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C24

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0525	0.0110
E & T	0.0535	0.0110
Good	0.0605	0.0150
Pooled	0.1420*	0.0620*
Satterthwaite	0.0530	0.0110

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C25

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0485	0.0130
E & T	0.0480	0.0125
Good	0.0585	0.0165
Pooled	0.2175*	0.1110*
Satterthwaite	0.0485	0.0130

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C26

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0430	0.0455	0.0505	0.0420	0.0505	0.0525	0.0485
	0.5	0.9985	0.9835	0.9335	0.7740	0.5325	0.3275	0.1295
	1.0	1.0000	1.0000	1.0000	0.9995	0.9840	0.8410	0.3290
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9930	0.6400
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8720
Efron & Tibshirani	0.0	0.0455	0.0490	0.0535	0.0455	0.0500	0.0535	0.0480
	0.5	0.9985	0.9840	0.9345	0.7770	0.5350	0.3285	0.1290
	1.0	1.0000	1.0000	1.0000	0.9995	0.9840	0.8420	0.3280
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9930	0.6385
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8685
Good	0.0	0.0505	0.0535	0.0585	0.0545	0.0595	0.0605	0.0585
	0.5	0.9985	0.9840	0.9385	0.7925	0.5635	0.3540	0.1550
	1.0	1.0000	1.0000	1.0000	1.0000	0.9905	0.8580	0.3610
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9935	0.6690
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8880
Pooled	0.0	0.0020	0.0060	0.0215	0.0450	0.0945	0.1420	0.2175
	0.5	0.9710	0.9260	0.8715	0.7810	0.6730	0.5485	0.3670
	1.0	1.0000	1.0000	1.0000	0.9995	0.9945	0.9355	0.6190
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8775
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9765
Satterthwaite	0.0	0.0455	0.0490	0.0530	0.0460	0.0505	0.0530	0.0485
	0.5	0.9985	0.9845	0.9345	0.7775	0.5390	0.3295	0.1295
	1.0	1.0000	1.0000	1.0000	0.9995	0.9850	0.8425	0.3295
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9930	0.6410
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8720

Table C27

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 120$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0085	0.0090	0.0095	0.0080	0.0085	0.0110	0.0130
	0.5	0.9935	0.9315	0.7850	0.5225	0.2705	0.1360	0.0325
	1.0	1.0000	1.0000	1.0000	0.9980	0.9245	0.6375	0.1490
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9620	0.3810
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6575
Efron & Tibshirani	0.0	0.0090	0.0105	0.0110	0.0090	0.0095	0.0110	0.0125
	0.5	0.9940	0.9380	0.7950	0.5325	0.2725	0.1360	0.0315
	1.0	1.0000	1.0000	1.0000	0.9975	0.9215	0.6330	0.1415
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9630	0.3710
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6480
Good	0.0	0.0105	0.0120	0.0125	0.0120	0.0145	0.0150	0.0165
	0.5	0.9945	0.9425	0.8200	0.5875	0.3120	0.1725	0.0510
	1.0	1.0000	1.0000	1.0000	0.9985	0.9420	0.6845	0.1830
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9710	0.4425
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.7230
Pooled	0.0	<.0005	<.0005	0.0025	0.0090	0.0260	0.0620	0.1110
	0.5	0.7855	0.7415	0.6460	0.5570	0.4550	0.3555	0.2405
	1.0	1.0000	1.0000	1.0000	0.9980	0.9750	0.8630	0.4660
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9950	0.7830
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9380
Satterthwaite	0.0	0.0095	0.0105	0.0105	0.0090	0.0100	0.0110	0.0130
	0.5	0.9940	0.9375	0.7975	0.5360	0.2775	0.1395	0.0325
	1.0	1.0000	1.0000	1.0000	0.9980	0.9285	0.6415	0.1495
	1.5	1.0000	1.0000	1.0000	1.0000	0.9995	0.9635	0.3830
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6580

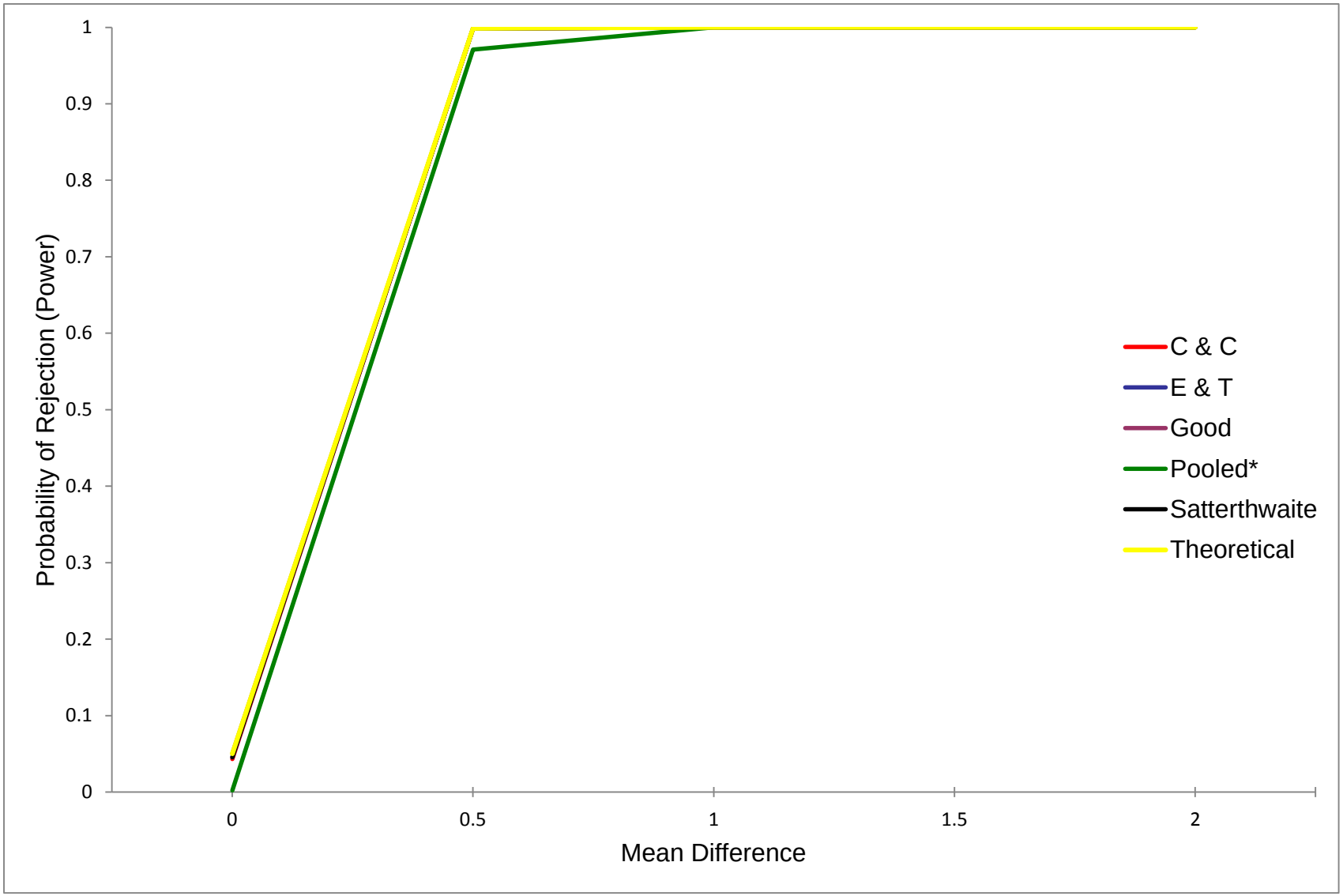

Figure C29. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

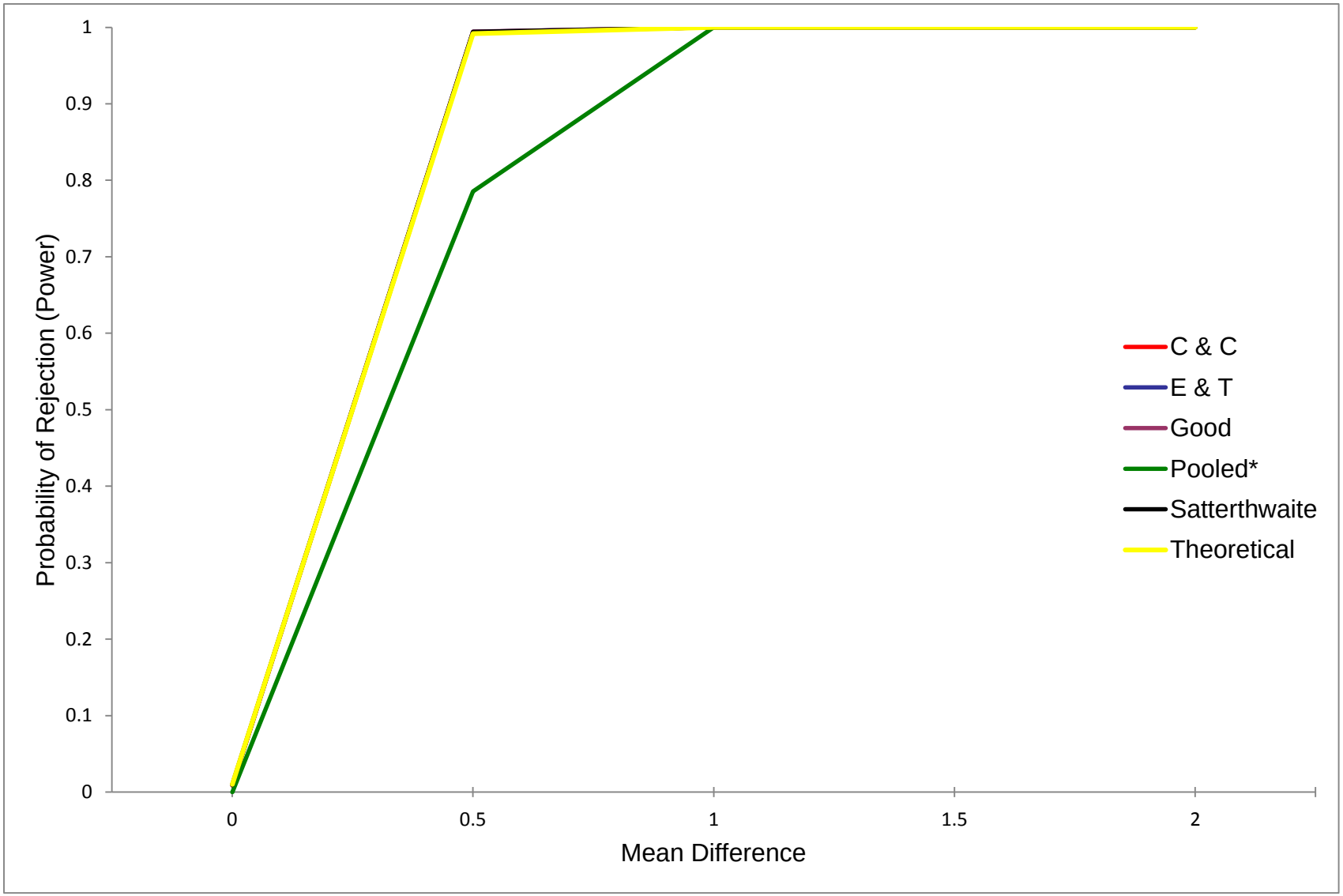

Figure C30. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

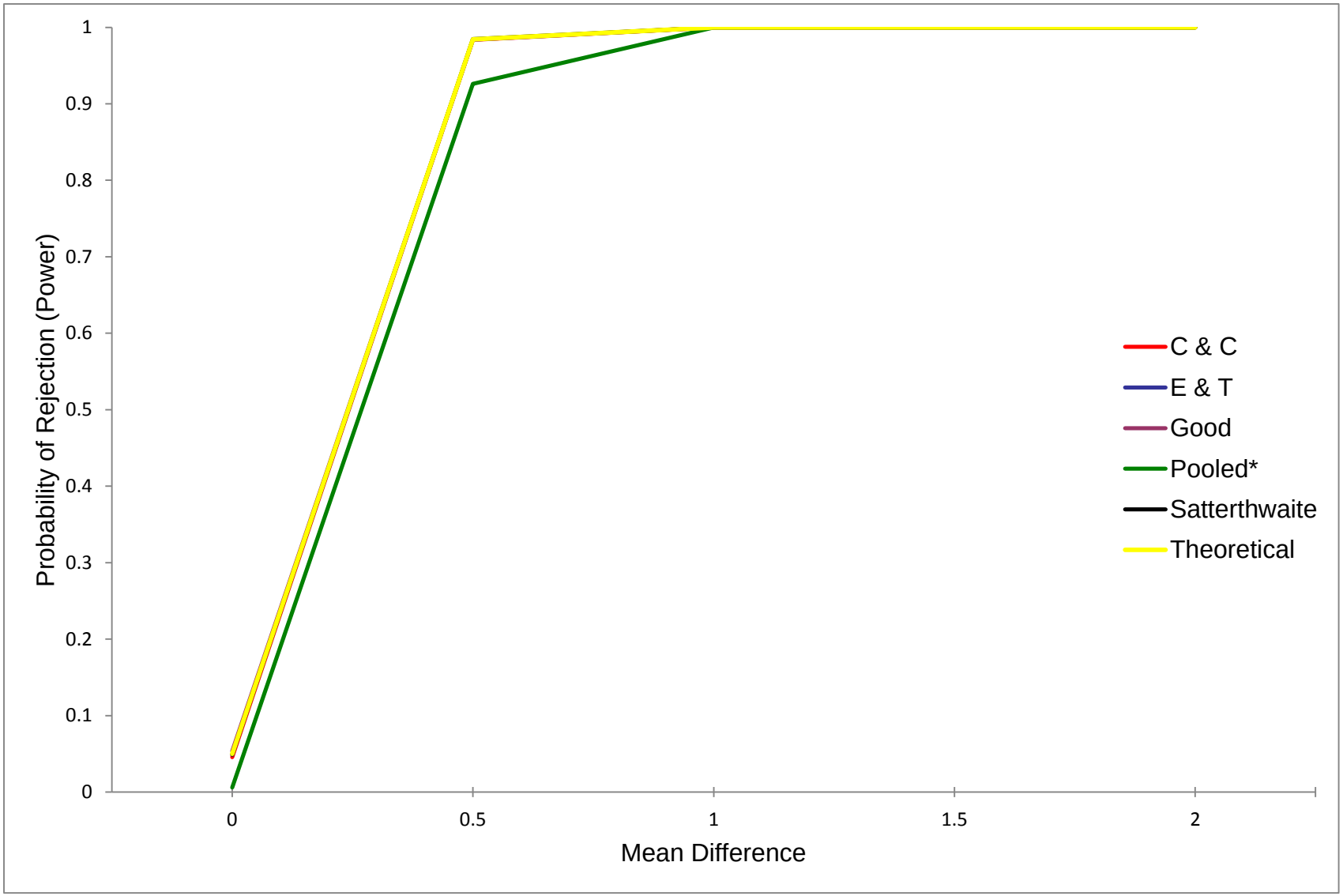

Figure C31. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

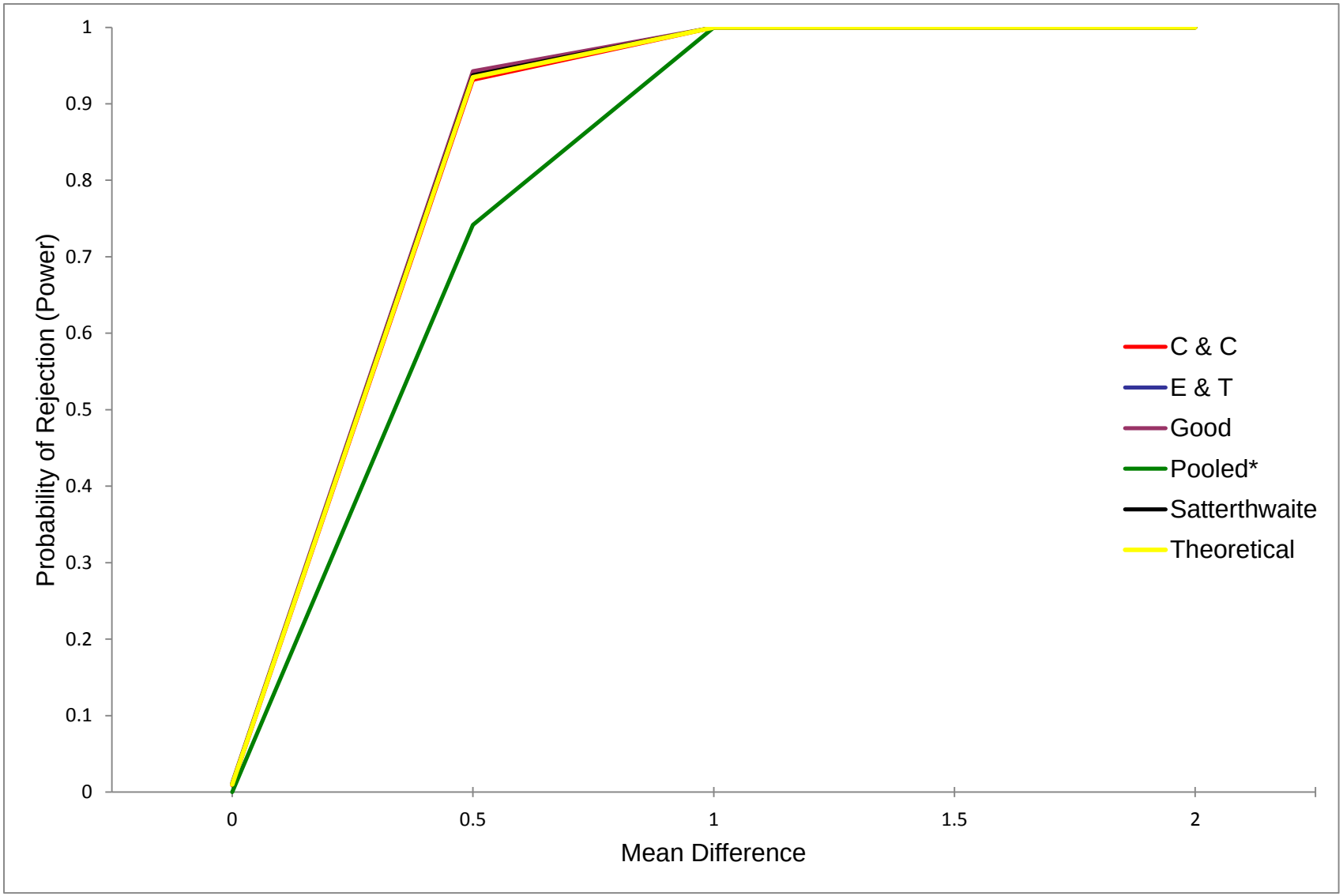

Figure C32. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

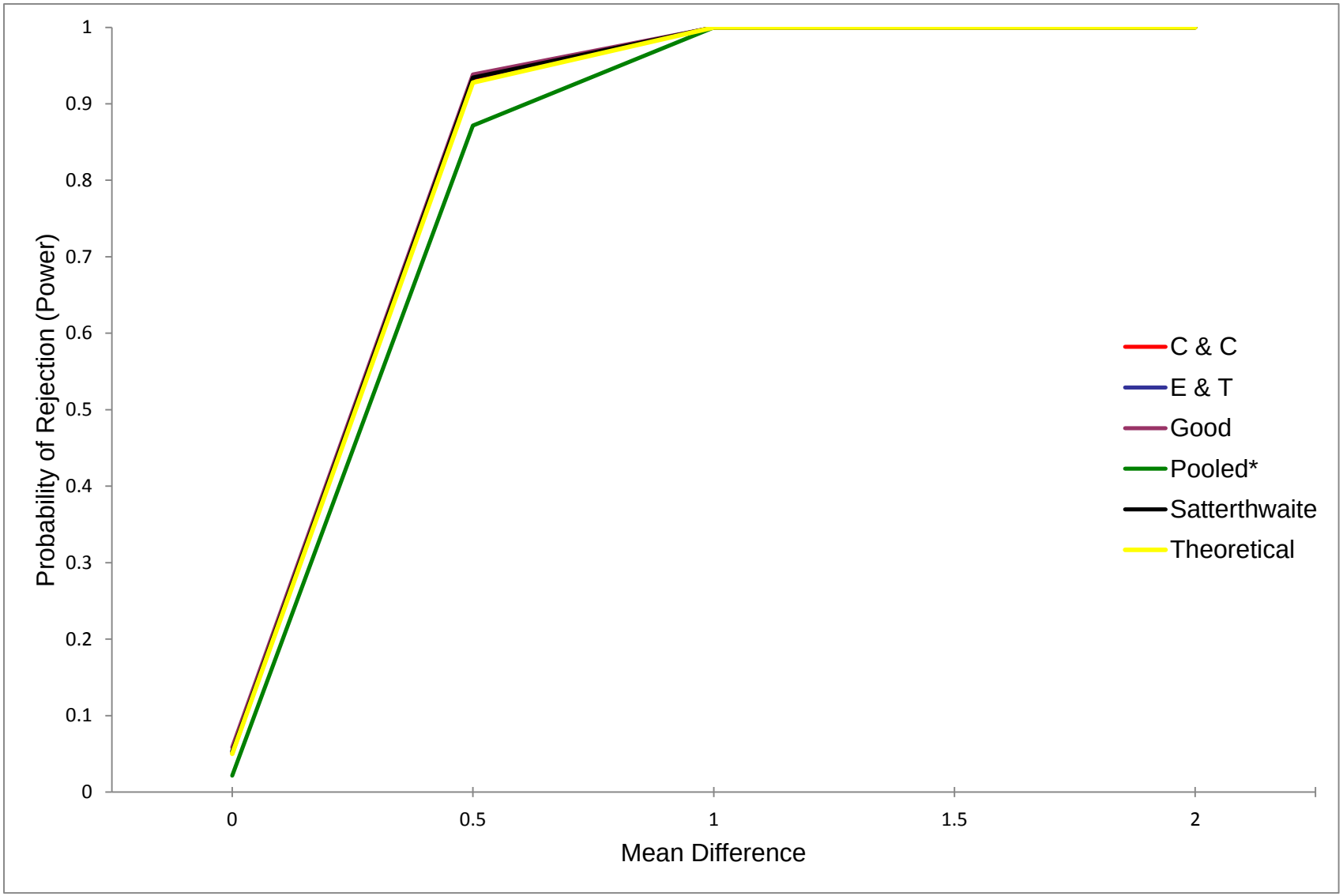

Figure C33. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

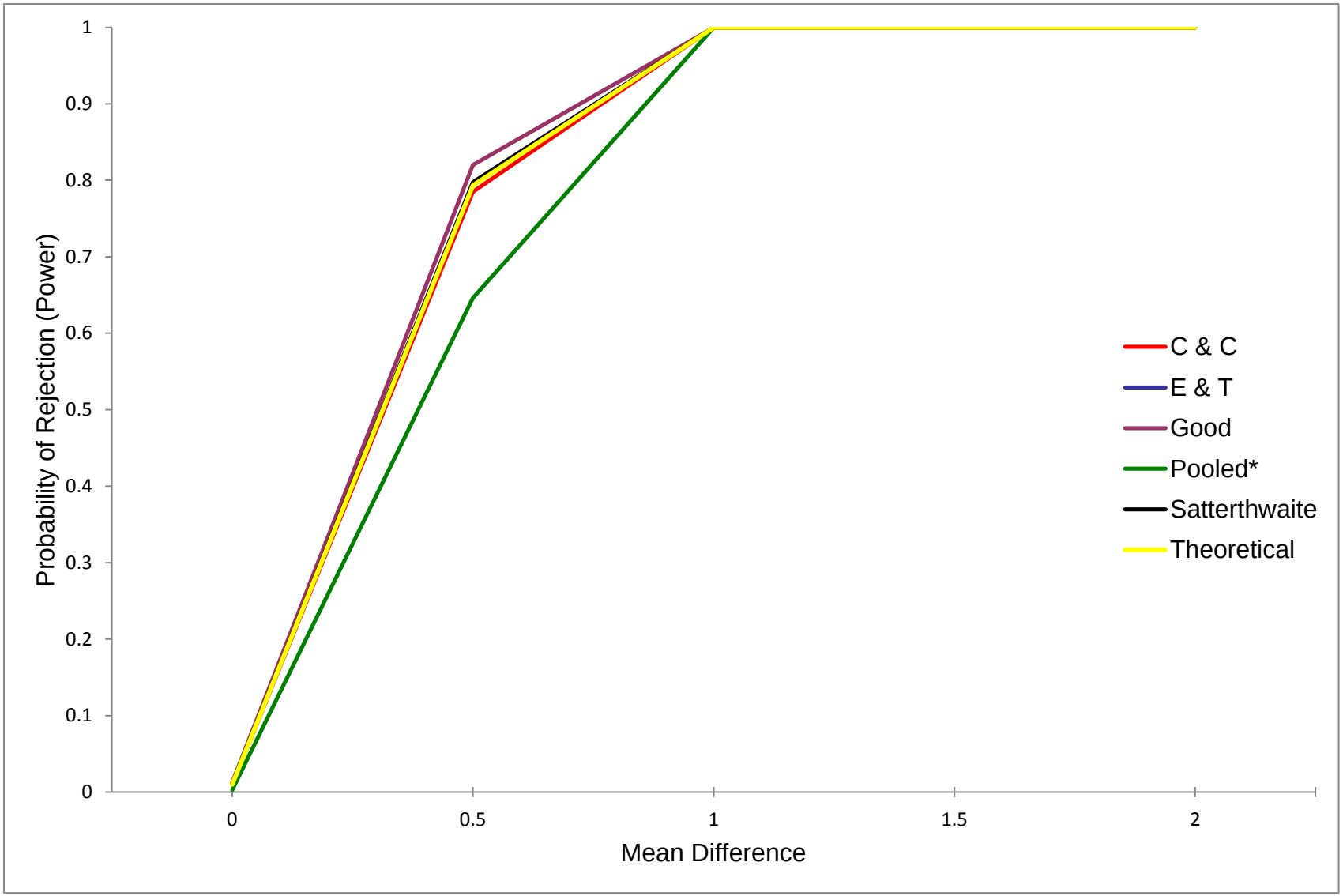

Figure C34. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

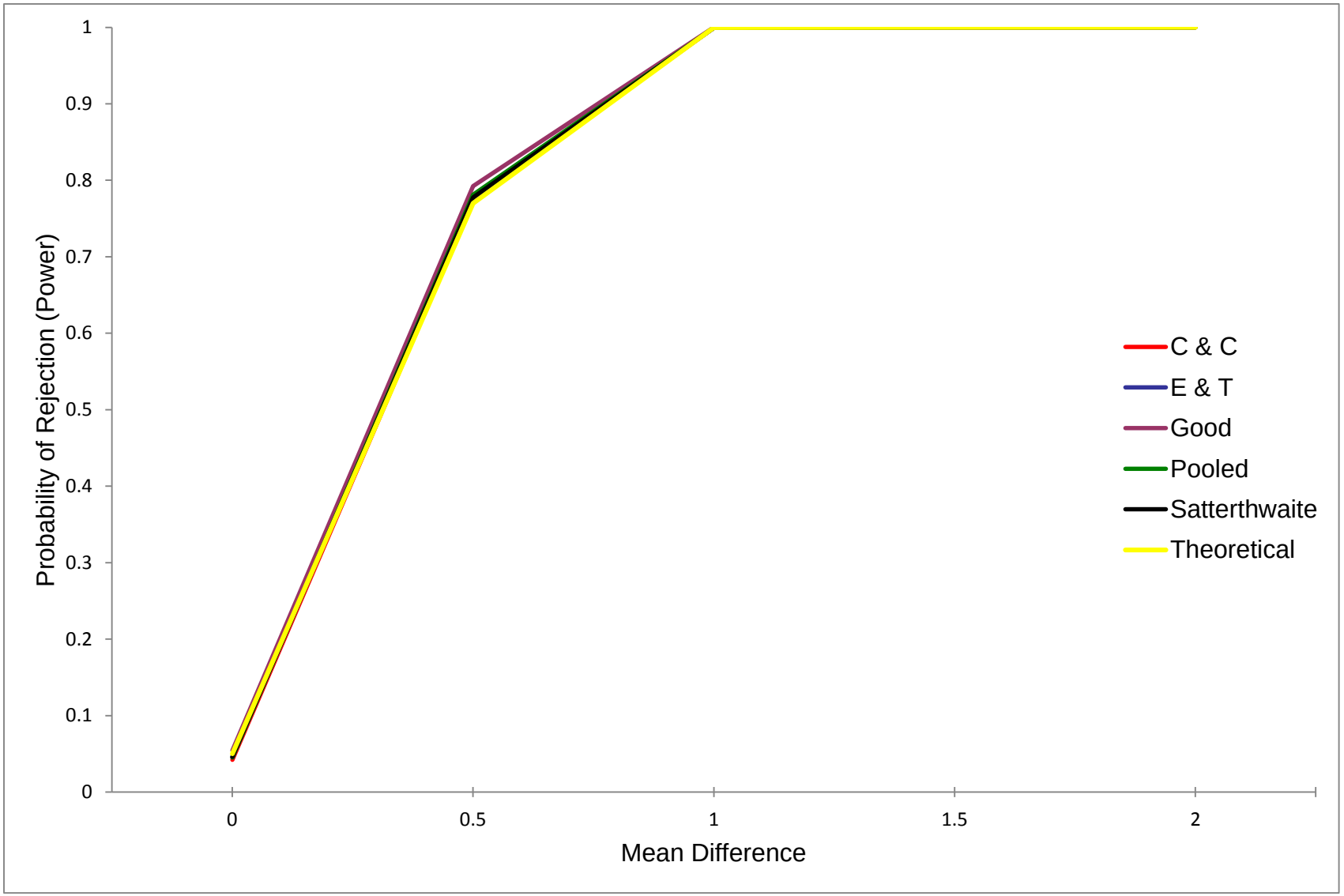

Figure C35. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

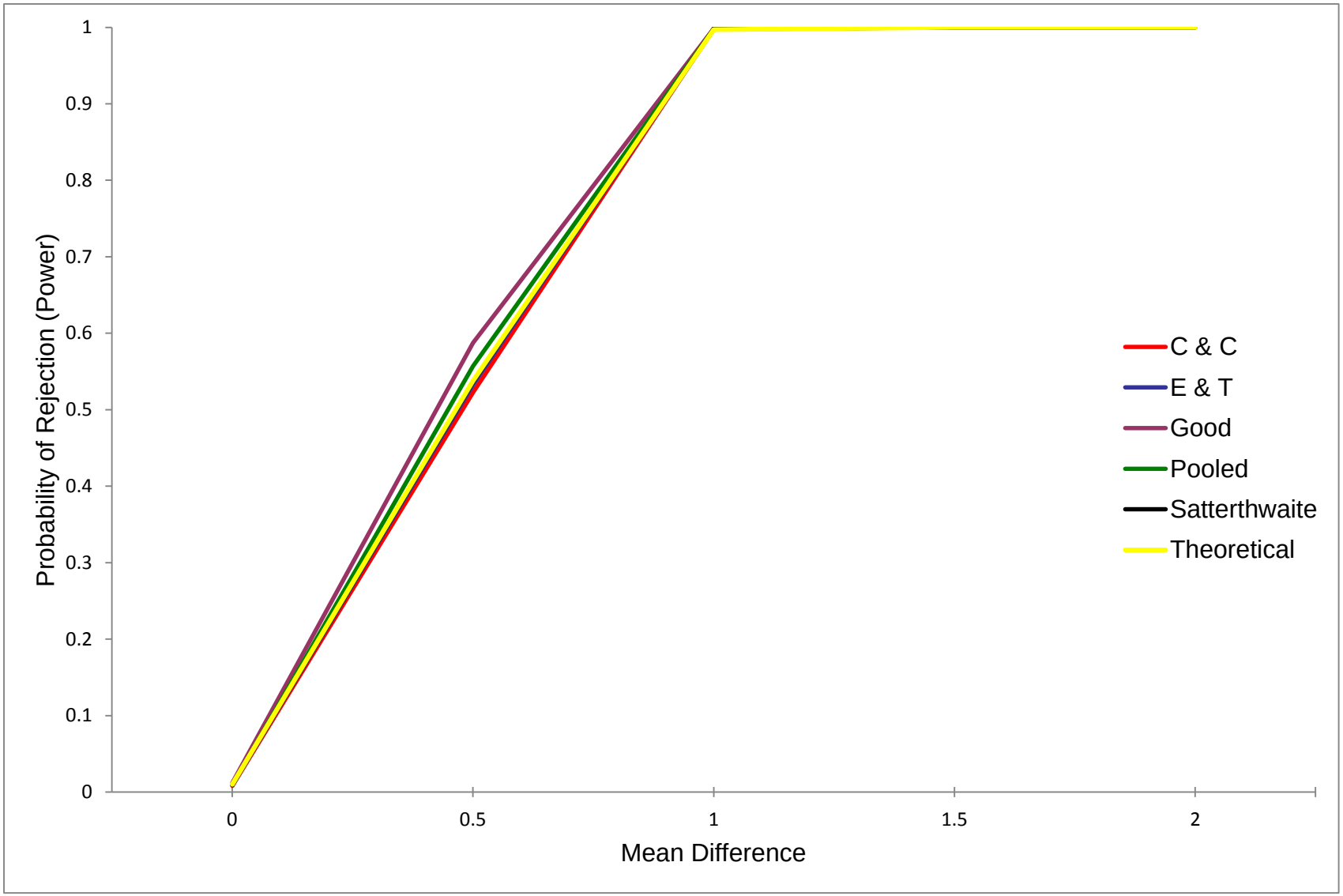

Figure C36. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

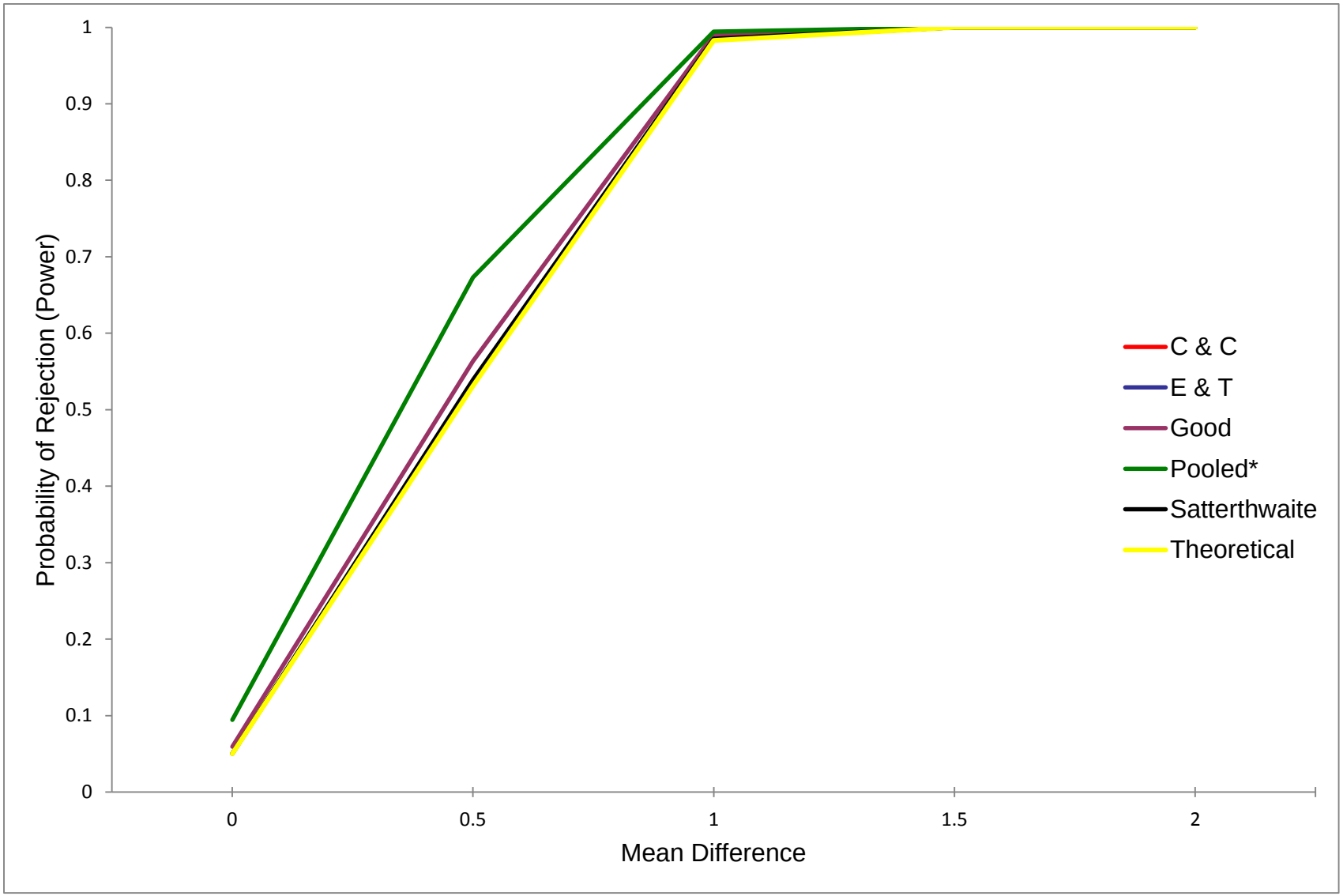

Figure C37. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

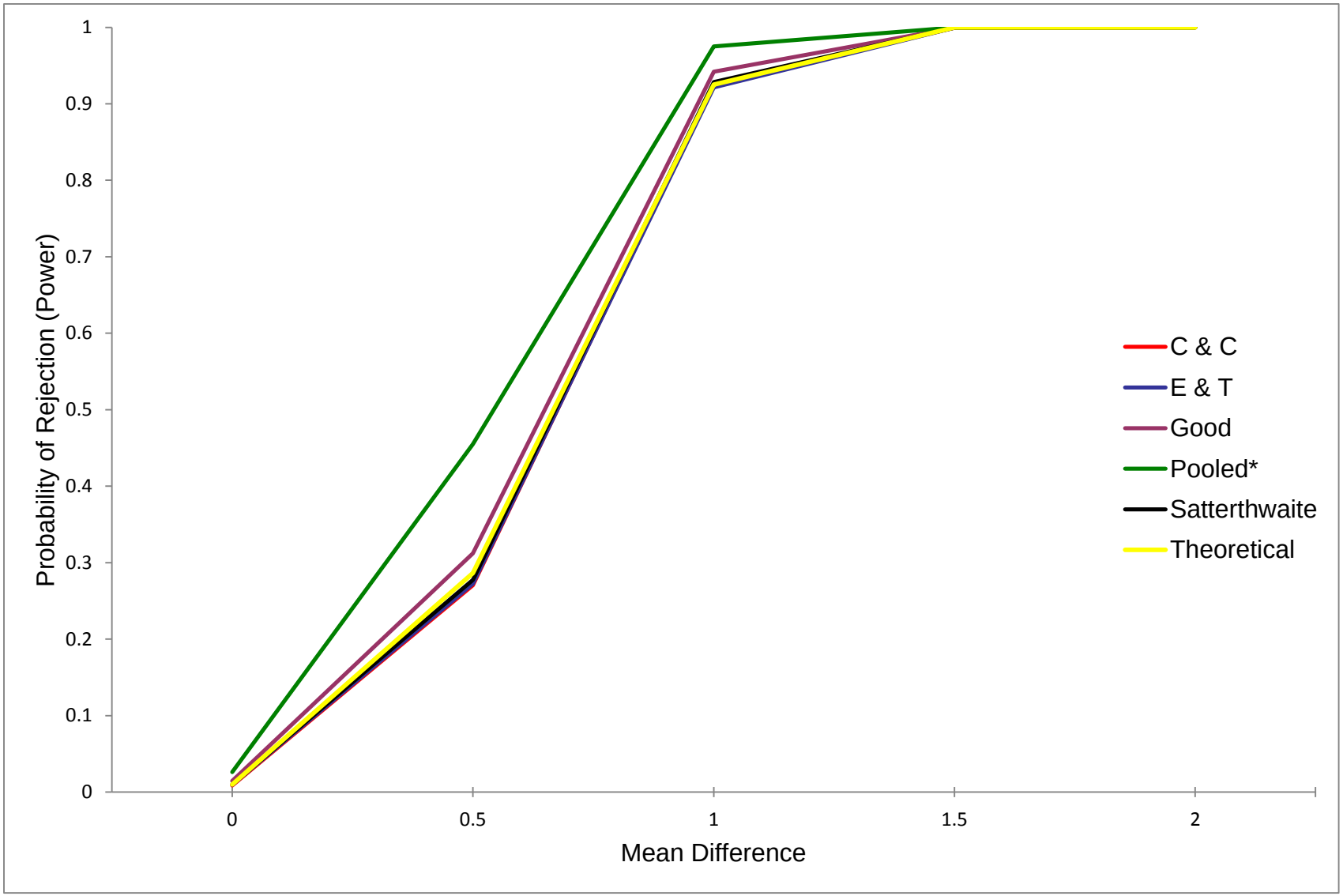

Figure C38. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

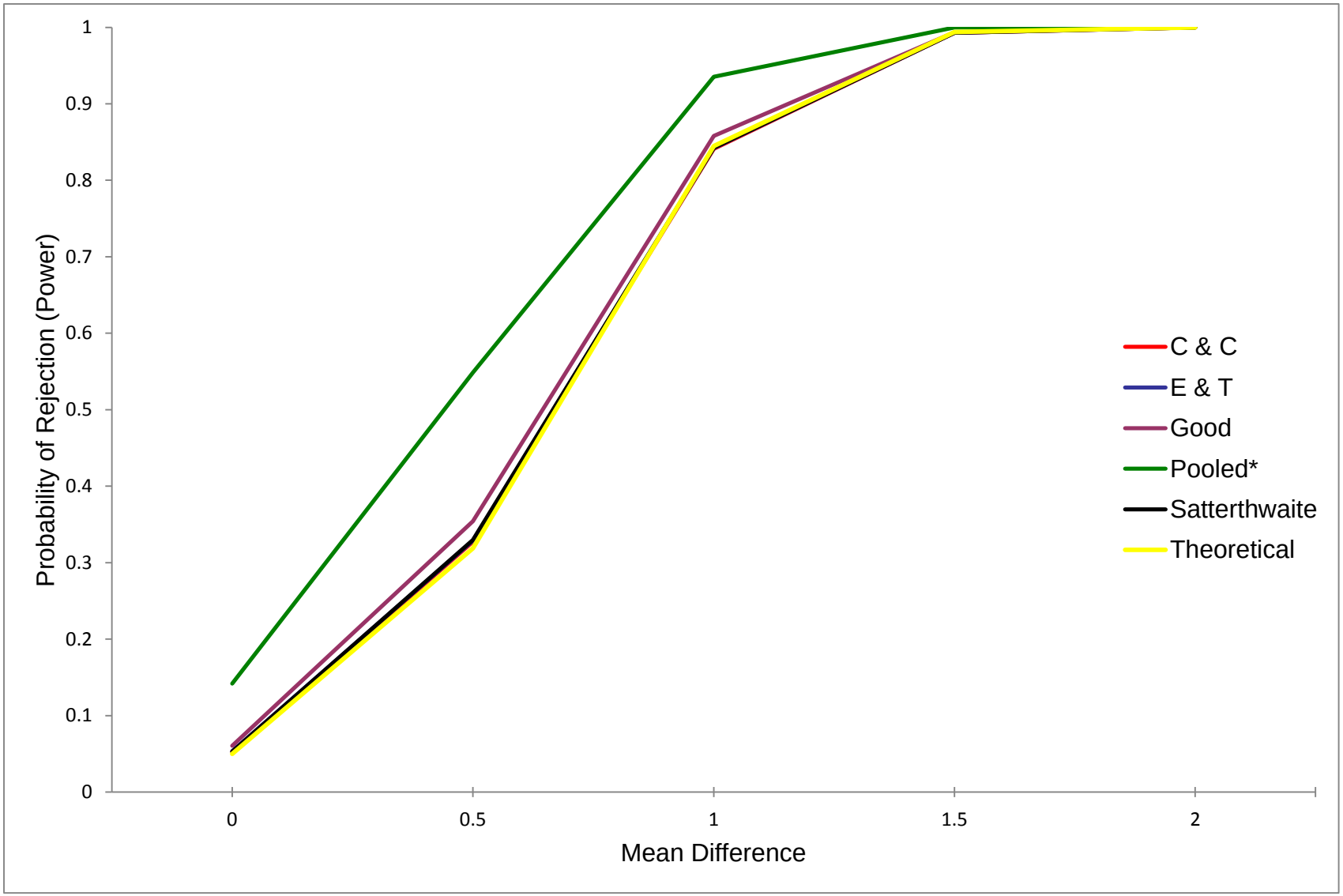

Figure C39. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

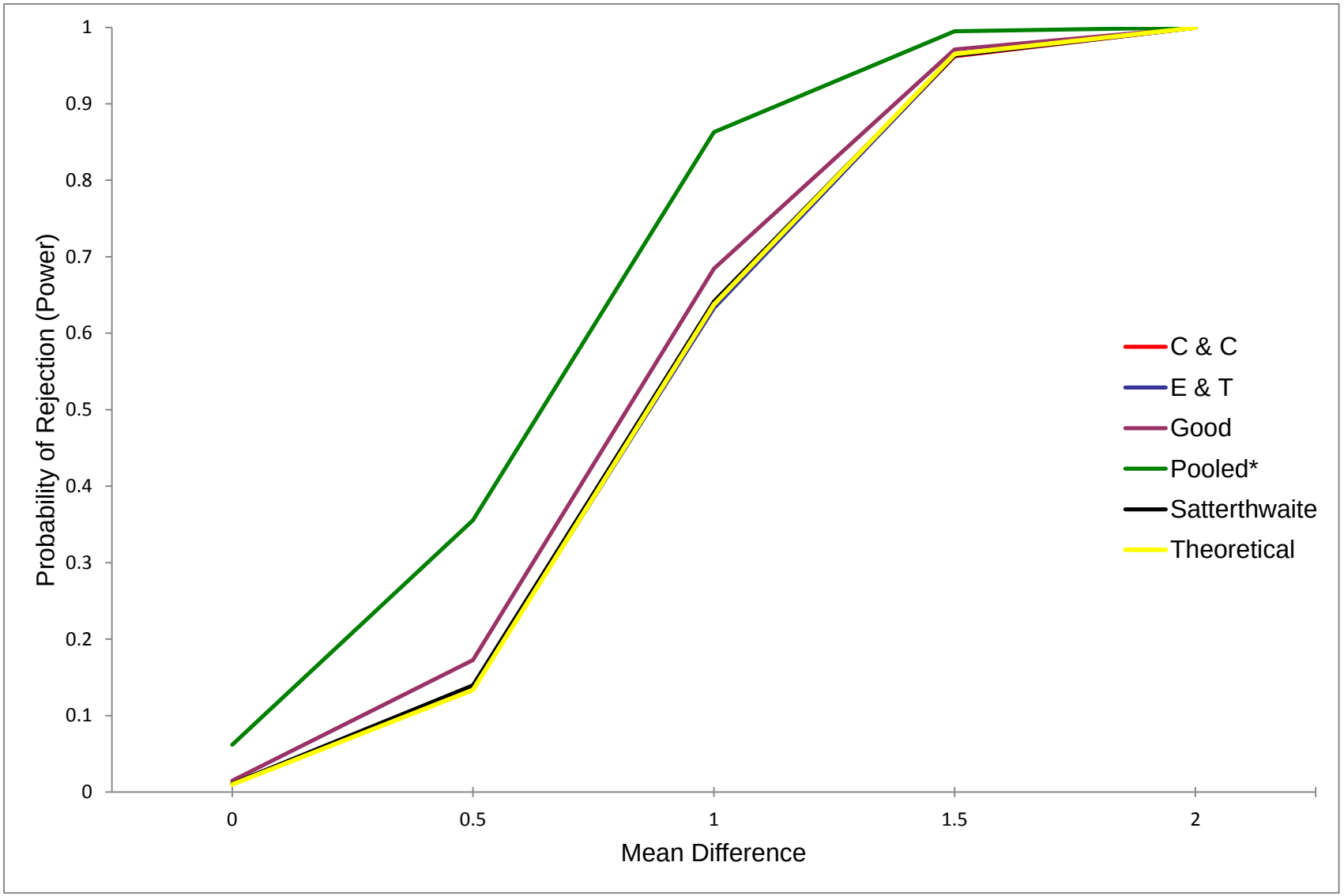

Figure C40. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

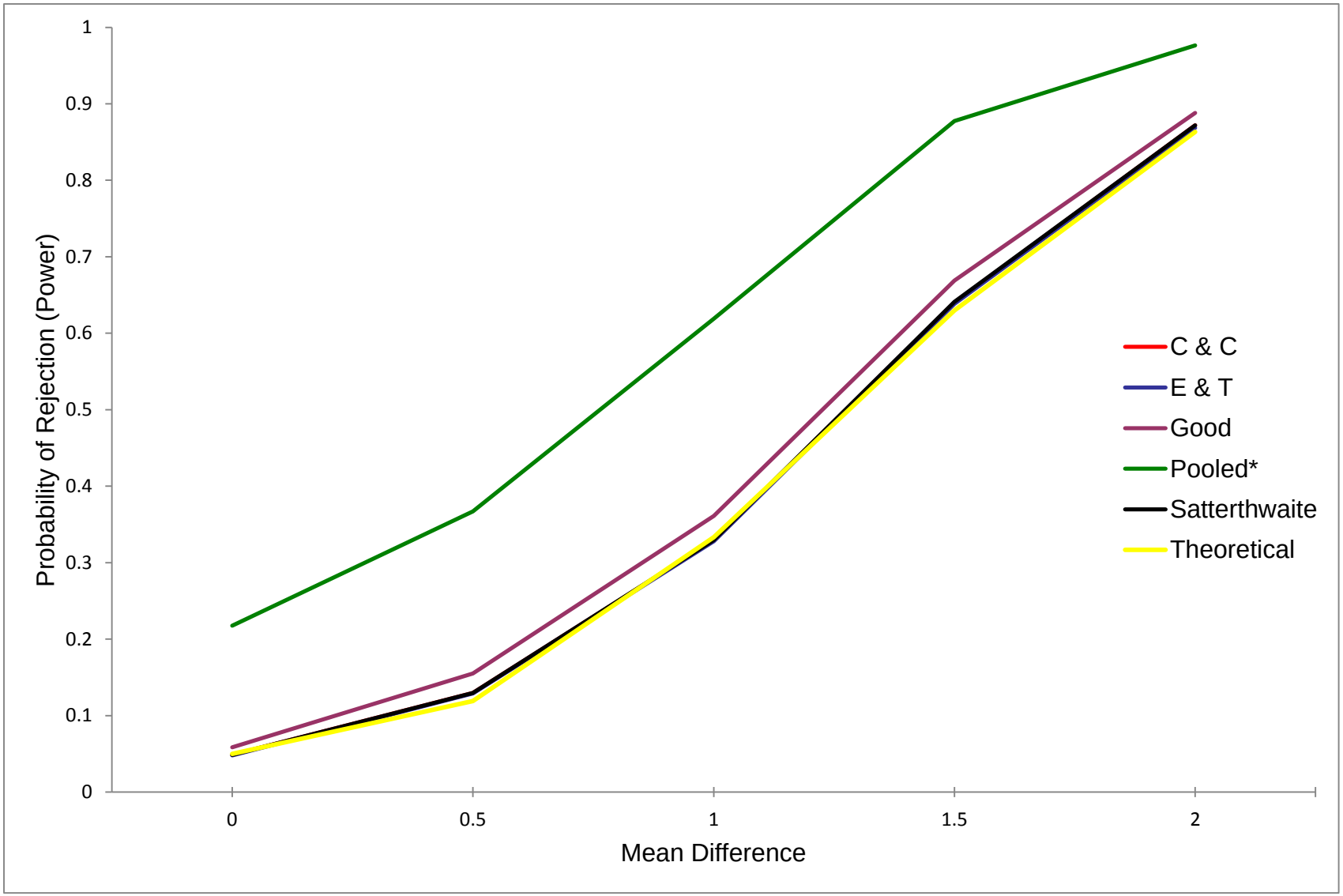

Figure C41. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

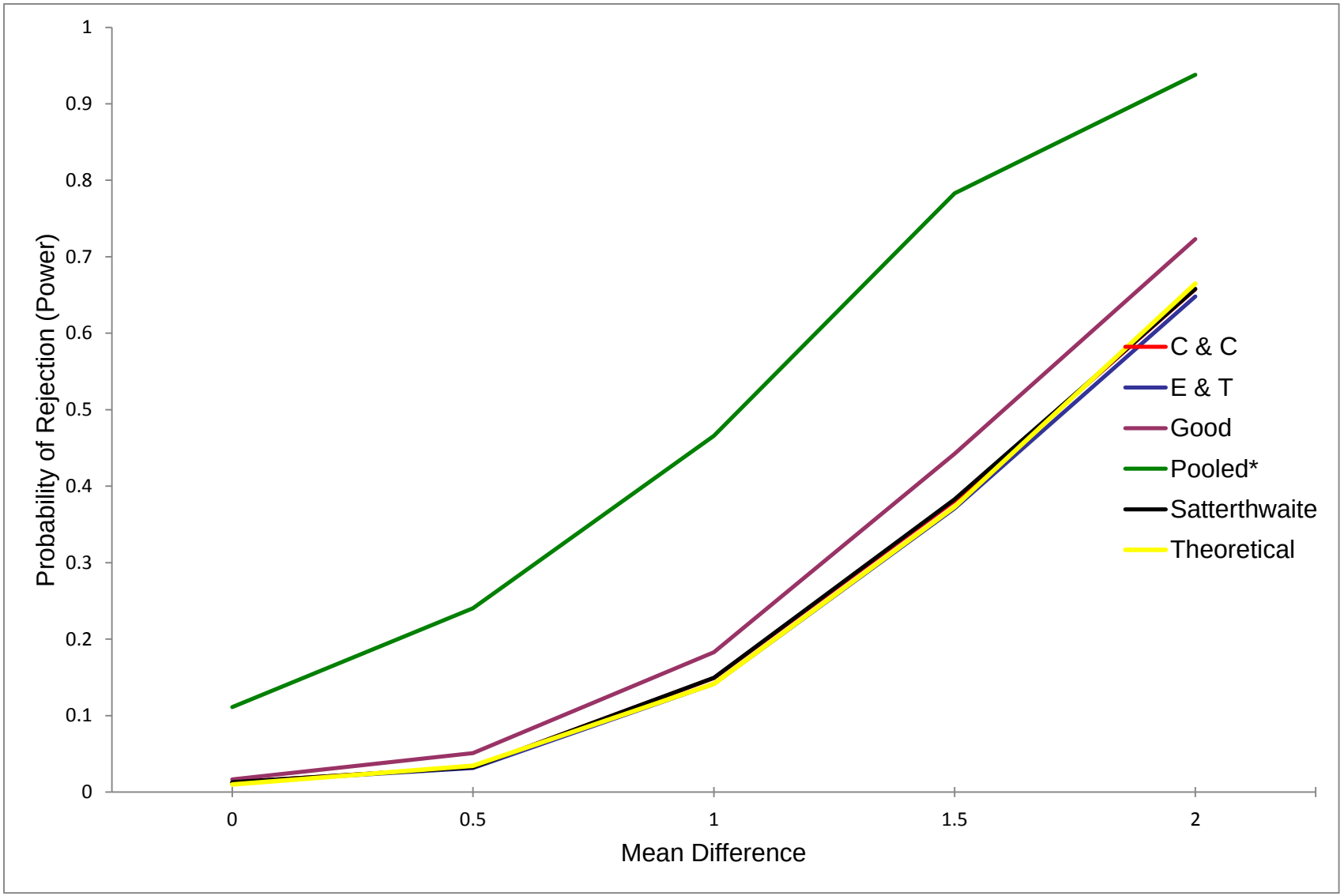

Figure C42. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 120$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

Sample-size Ratio (n_2/n_1) was 5.0 (i.e., $n_1 = 40$, $n_2 = 200$)

Table C28

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0495	0.0060
E & T	0.0505	0.0060
Good	0.0530	0.0085
Pooled	<.0005*	<.0005*
Satterthwaite	0.0510	0.0065

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C29

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1/4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0425	0.0090
E & T	0.0425	0.0095
Good	0.0455	0.0110
Pooled	0.0010*	<.0005*
Satterthwaite	0.0430	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C30

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) *was 1/2, at two standards (i.e., .01 and .05)*

Method	α = .05	α = .01
C & C	0.0505	0.0090
E & T	0.0510	0.0095
Good	0.0595	0.0110
Pooled	0.0130*	0.0010*
Satterthwaite	0.0515	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C31

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 1, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0475	0.0095
E & T	0.0485	0.0095
Good	0.0540	0.0120
Pooled	0.0445	0.0080
Satterthwaite	0.0490	0.0100

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C32

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 2, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0480	0.0060
E & T	0.0475	0.0060
Good	0.0575	0.0105
Pooled	0.1245*	0.0350*
Satterthwaite	0.0490	0.0060

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C33

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 4, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0545	0.0095
E & T	0.0560	0.0090
Good	0.0635	0.0165
Pooled	0.2060*	0.0945*
Satterthwaite	0.0555	0.0095

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C34

Type I error rates of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, and variance ratio (var_1/var_2) was 16, at two standards (i.e., .01 and .05)

Method	α = .05	α = .01
C & C	0.0525	0.0140
E & T	0.0510	0.0130
Good	0.0590	0.0195*
Pooled	0.3280*	0.1980*
Satterthwaite	0.0525	0.0140

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.
* $p \leq .001$ (significant at α = .001).

Table C35

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .05

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0495	0.0425	0.0505	0.0475	0.0480	0.0545	0.0525
	0.5	1.0000	0.9970	0.9650	0.8030	0.5450	0.3245	0.1205
	1.0	1.0000	1.0000	1.0000	1.0000	0.9880	0.8515	0.3285
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9960	0.6100
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8610
Efron & Tibshirani	0.0	0.0505	0.0425	0.0510	0.0485	0.0475	0.0560	0.0510
	0.5	1.0000	0.9970	0.9675	0.8025	0.5495	0.3250	0.1180
	1.0	1.0000	1.0000	1.0000	1.0000	0.9880	0.8500	0.3275
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9955	0.6090
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8605
Good	0.0	0.0530	0.0455	0.0595	0.0540	0.0575	0.0635	0.0590
	0.5	1.0000	0.9970	0.9710	0.8235	0.5790	0.3565	0.1380
	1.0	1.0000	1.0000	1.0000	1.0000	0.9895	0.8655	0.3610
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9970	0.6420
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8795
Pooled	0.0	<.0005	0.0010	0.0130	0.0445	0.1245	0.2060	0.3280
	0.5	0.9900	0.9570	0.9035	0.8205	0.7090	0.6080	0.4570
	1.0	1.0000	1.0000	1.0000	1.0000	0.9985	0.9635	0.7210
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9030
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9835
Satterthwaite	0.0	0.0510	0.0430	0.0515	0.0490	0.0490	0.0555	0.0525
	0.5	1.0000	0.9970	0.9675	0.8045	0.5490	0.3250	0.1205
	1.0	1.0000	1.0000	1.0000	1.0000	0.9880	0.8515	0.3295
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9960	0.6105
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8615

Table C36

Power values of all methods based on 2,000 replications of the simulated experiment when $n_1 = 40$, $n_2 = 200$, var_2 was fixed to 1, variance ratio (VR) is equal to (var_1/var_2), at a standard = .01

Method	Mean Difference	VR = 1/16 = 0.0625	VR = 1/4 = 0.25	VR = 1/2 = 0.5	VR = 1	VR = 2	VR = 4	VR = 16
Cochran & Cox	0.0	0.0060	0.0090	0.0090	0.0095	0.0060	0.0095	0.0140
	0.5	1.0000	0.9840	0.8595	0.5785	0.2910	0.1205	0.0325
	1.0	1.0000	1.0000	1.0000	0.9960	0.9405	0.6335	0.1485
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9765	0.3655
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6665
Efron & Tibshirani	0.0	0.0060	0.0095	0.0095	0.0095	0.0060	0.0090	0.0130
	0.5	1.0000	0.9870	0.8670	0.5775	0.2895	0.1180	0.0320
	1.0	1.0000	1.0000	1.0000	0.9965	0.9370	0.6325	0.1460
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9755	0.3490
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6510
Good	0.0	0.0085	0.0110	0.0110	0.0120	0.0105	0.0165	0.0195
	0.5	1.0000	0.9895	0.8945	0.6345	0.3510	0.1600	0.0445
	1.0	1.0000	1.0000	1.0000	0.9985	0.9555	0.6865	0.1890
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9810	0.4255
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.7205
Pooled	0.0	<.0005	<.0005	0.0010	0.0080	0.0350	0.0945	0.1980
	0.5	0.8465	0.7715	0.7050	0.6060	0.5150	0.4440	0.3195
	1.0	1.0000	1.0000	1.0000	0.9980	0.9865	0.9135	0.5900
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9980	0.8300
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9670
Satterthwaite	0.0	0.0065	0.0095	0.0095	0.0100	0.0060	0.0095	0.0140
	0.5	1.0000	0.9860	0.8680	0.5860	0.2950	0.1230	0.0325
	1.0	1.0000	1.0000	1.0000	0.9970	0.9415	0.6375	0.1495
	1.5	1.0000	1.0000	1.0000	1.0000	1.0000	0.9770	0.3660
	2.0	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.6665

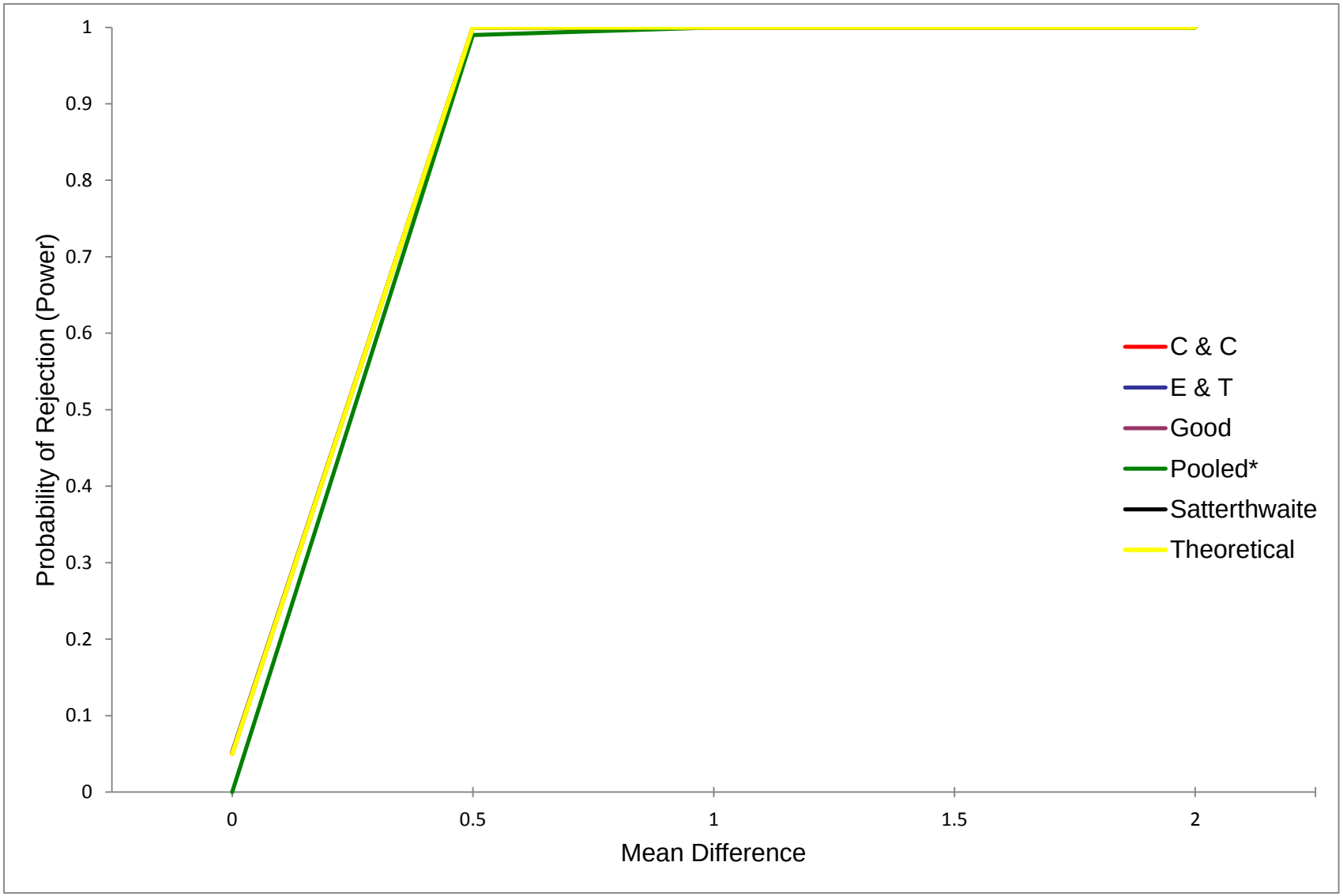

Figure C43. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

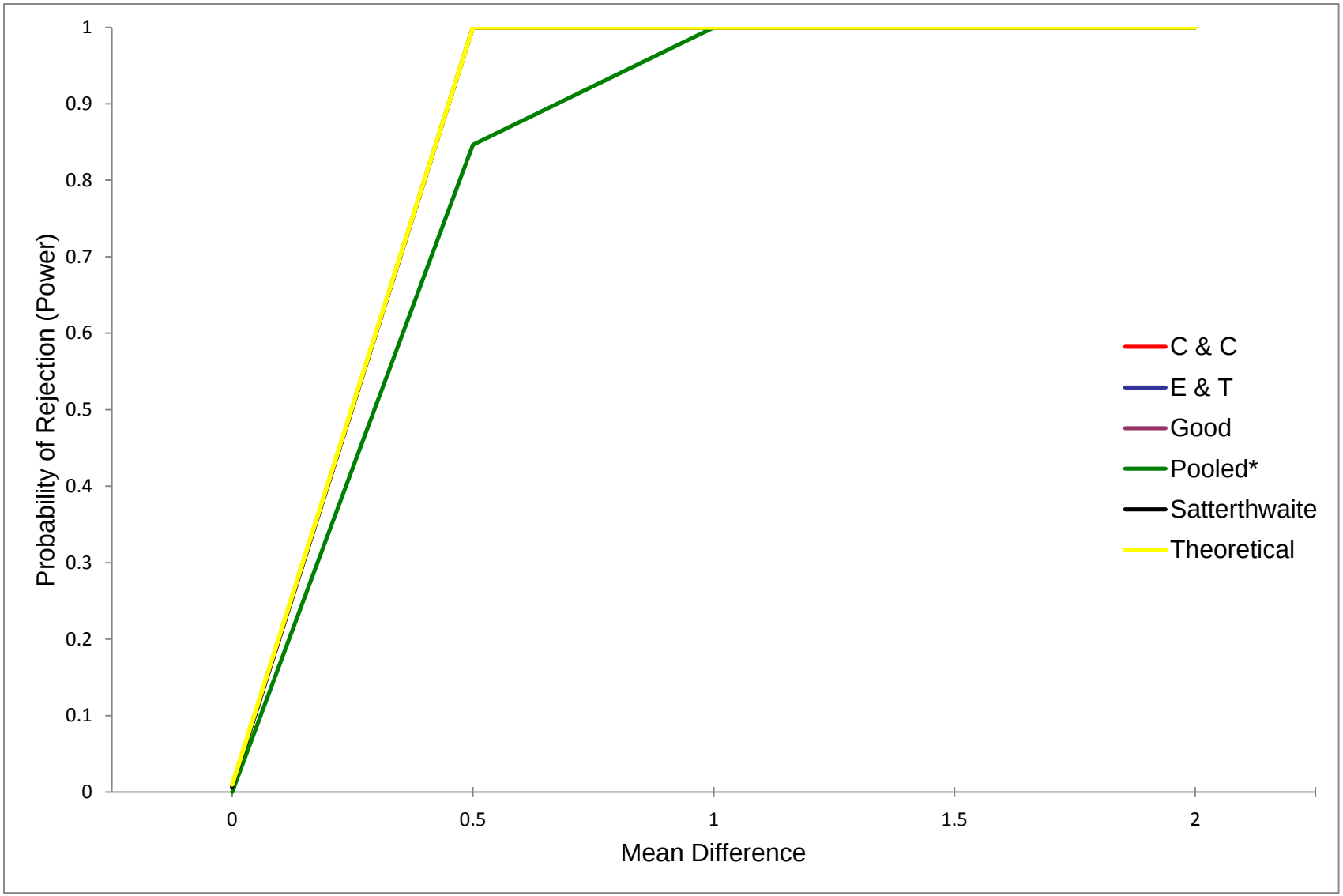

Figure C44. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

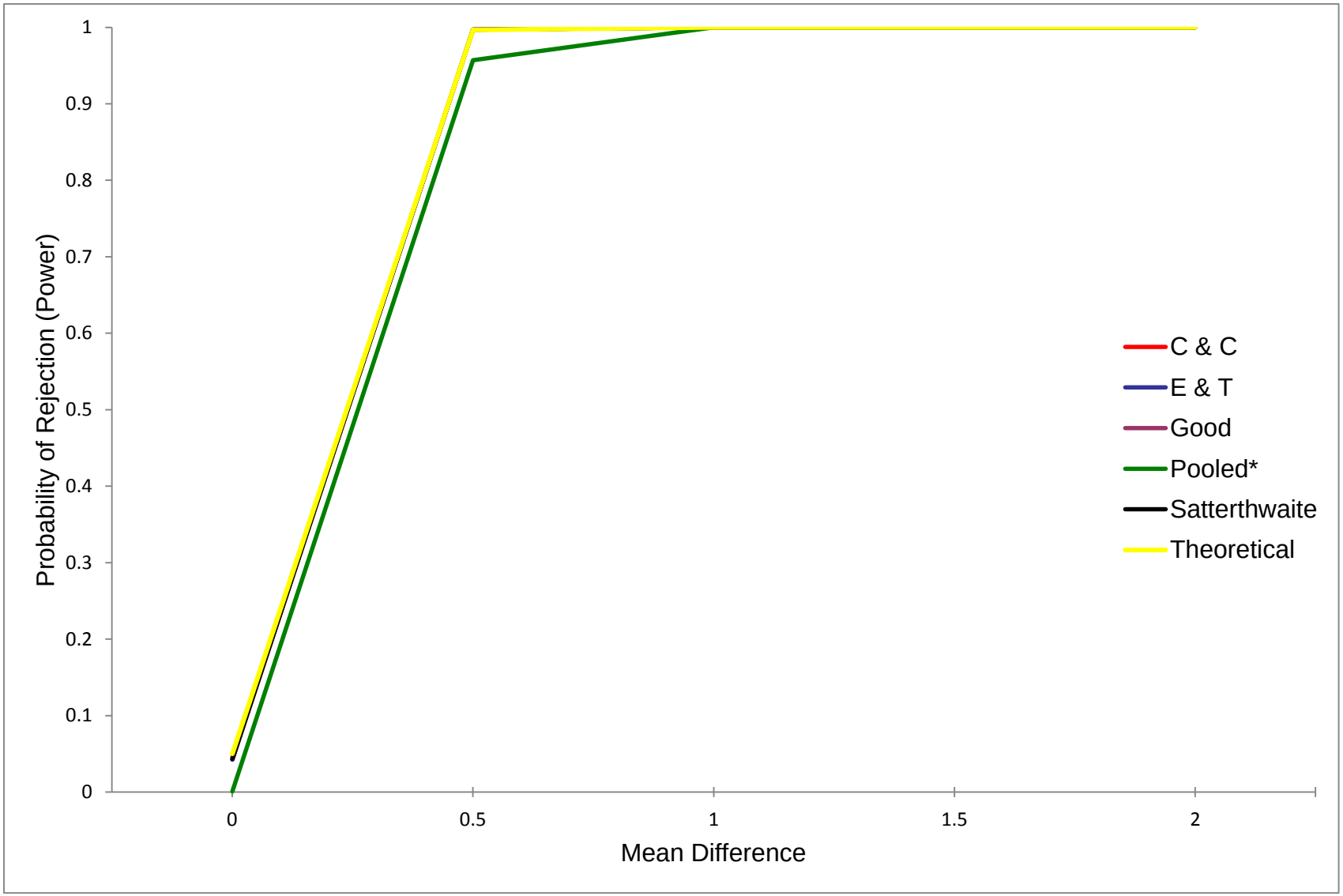

Figure C45. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

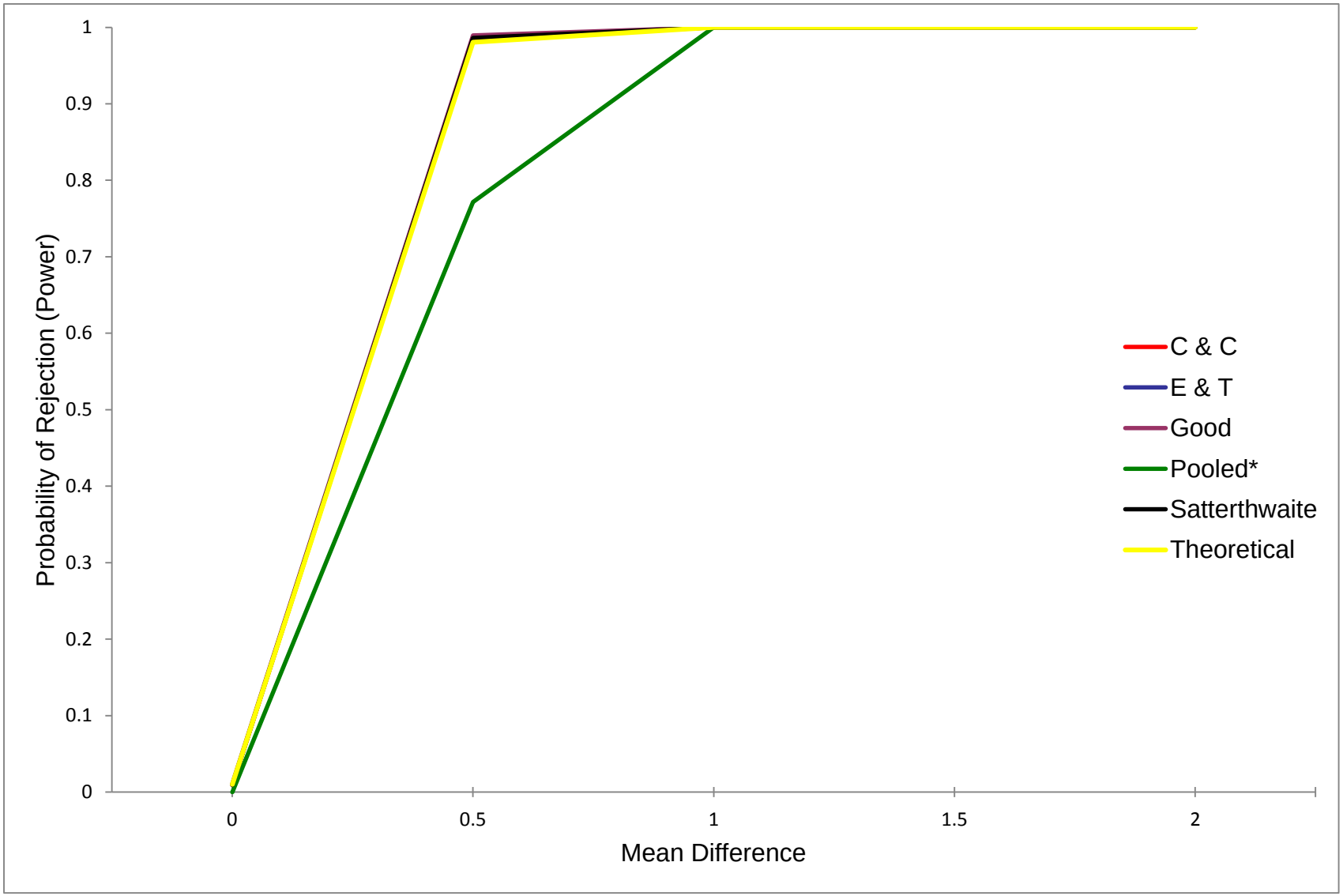

Figure C46. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

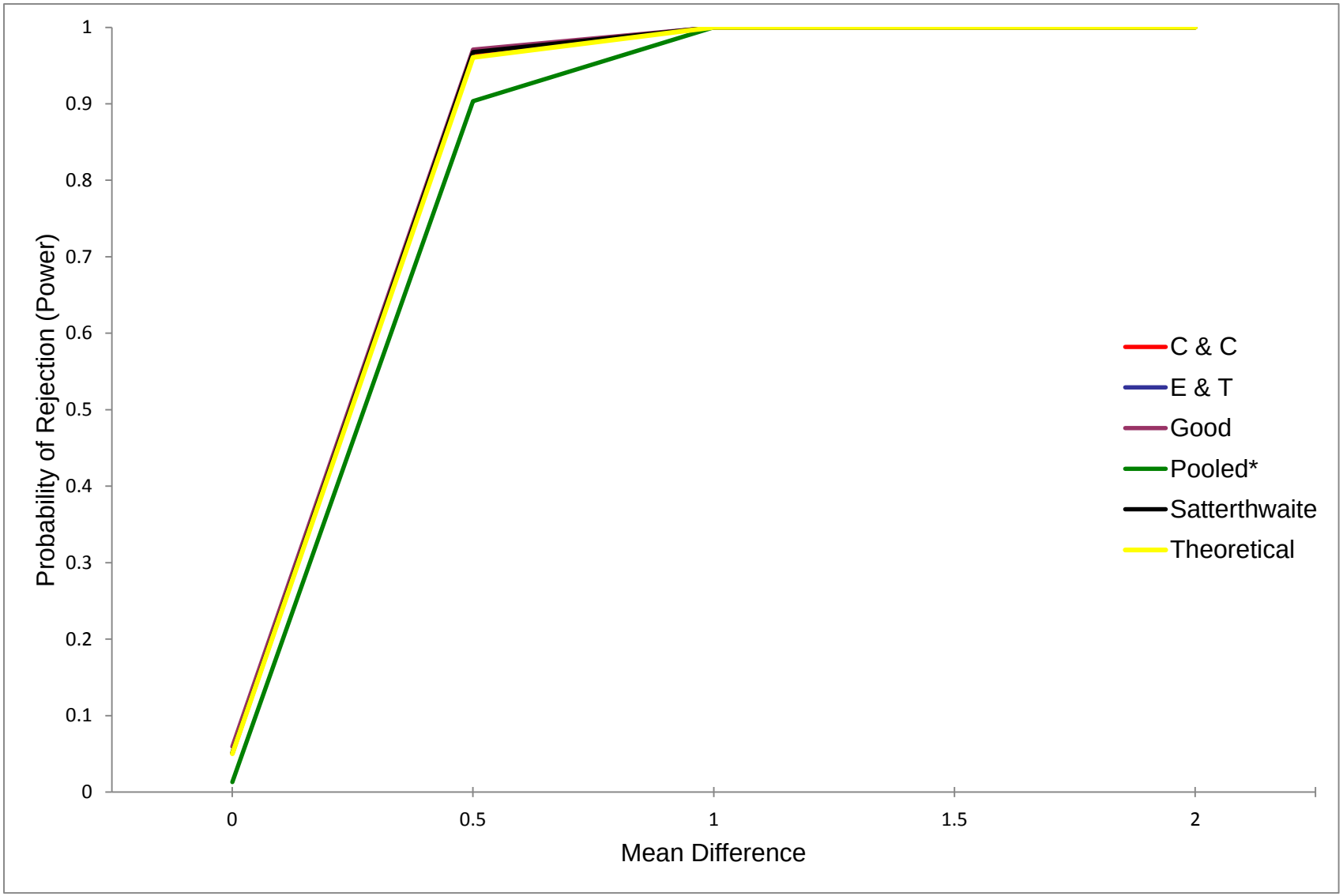

Figure C47. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

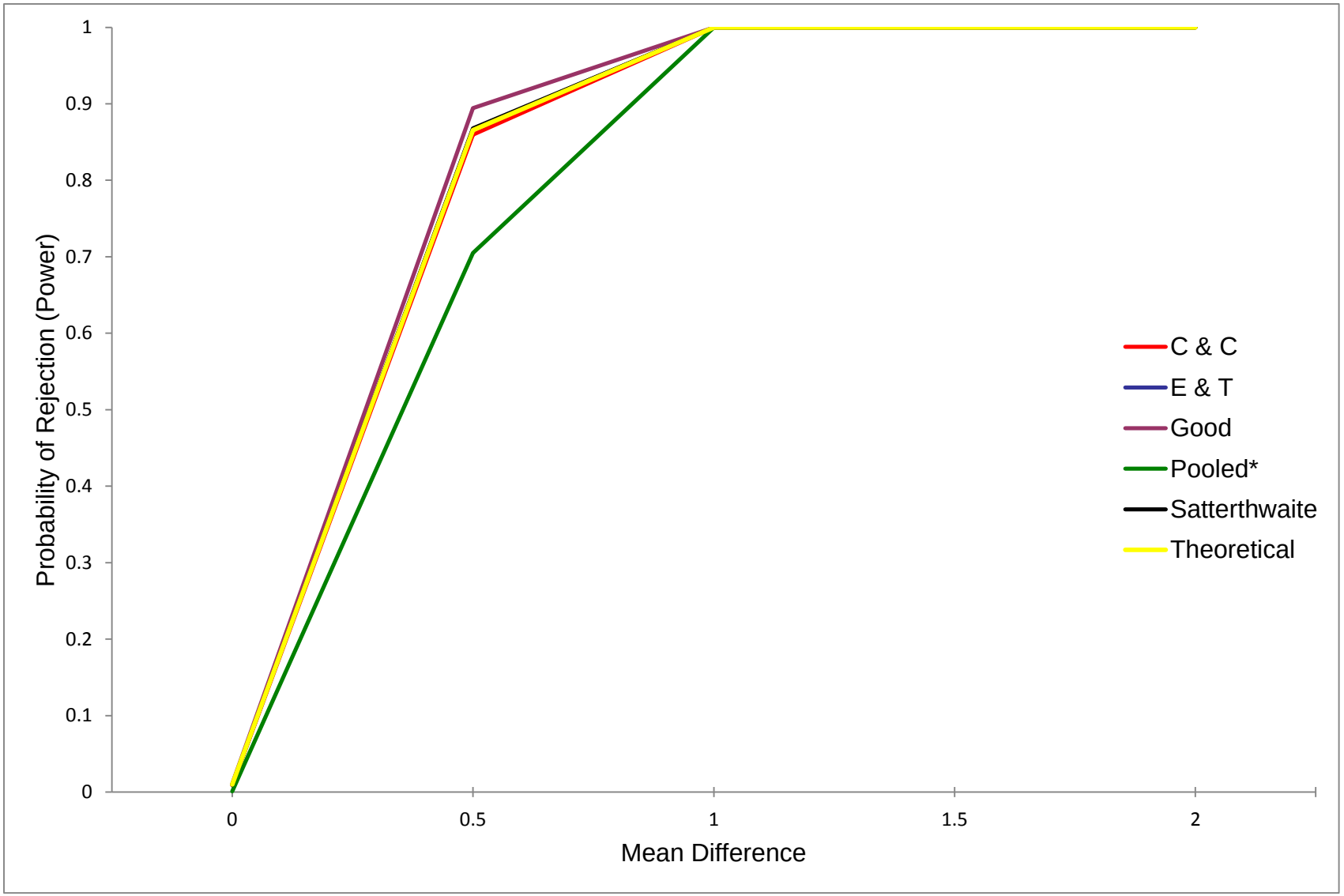

Figure C48. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1/2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

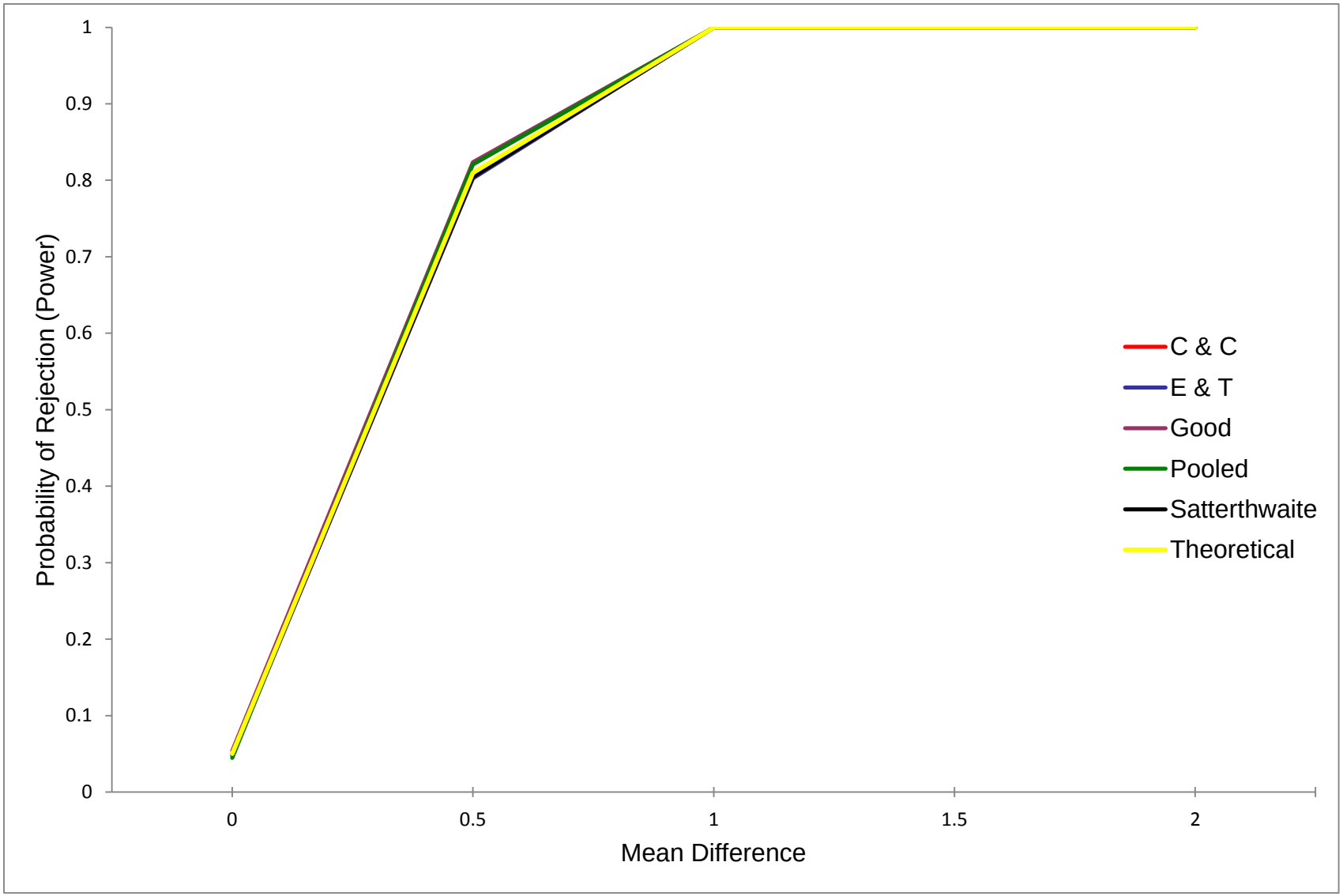

Figure C49. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

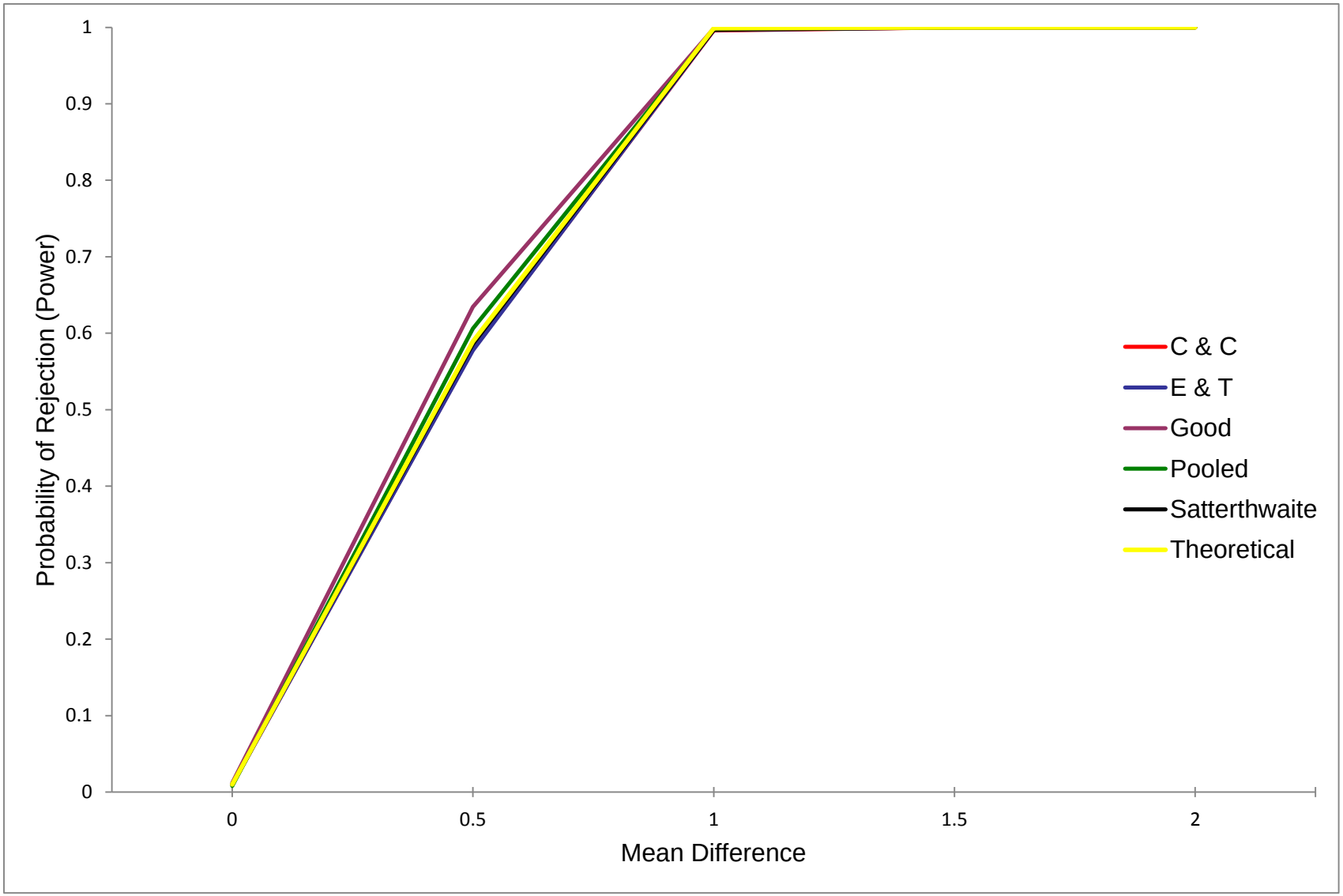

Figure C50. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 1, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

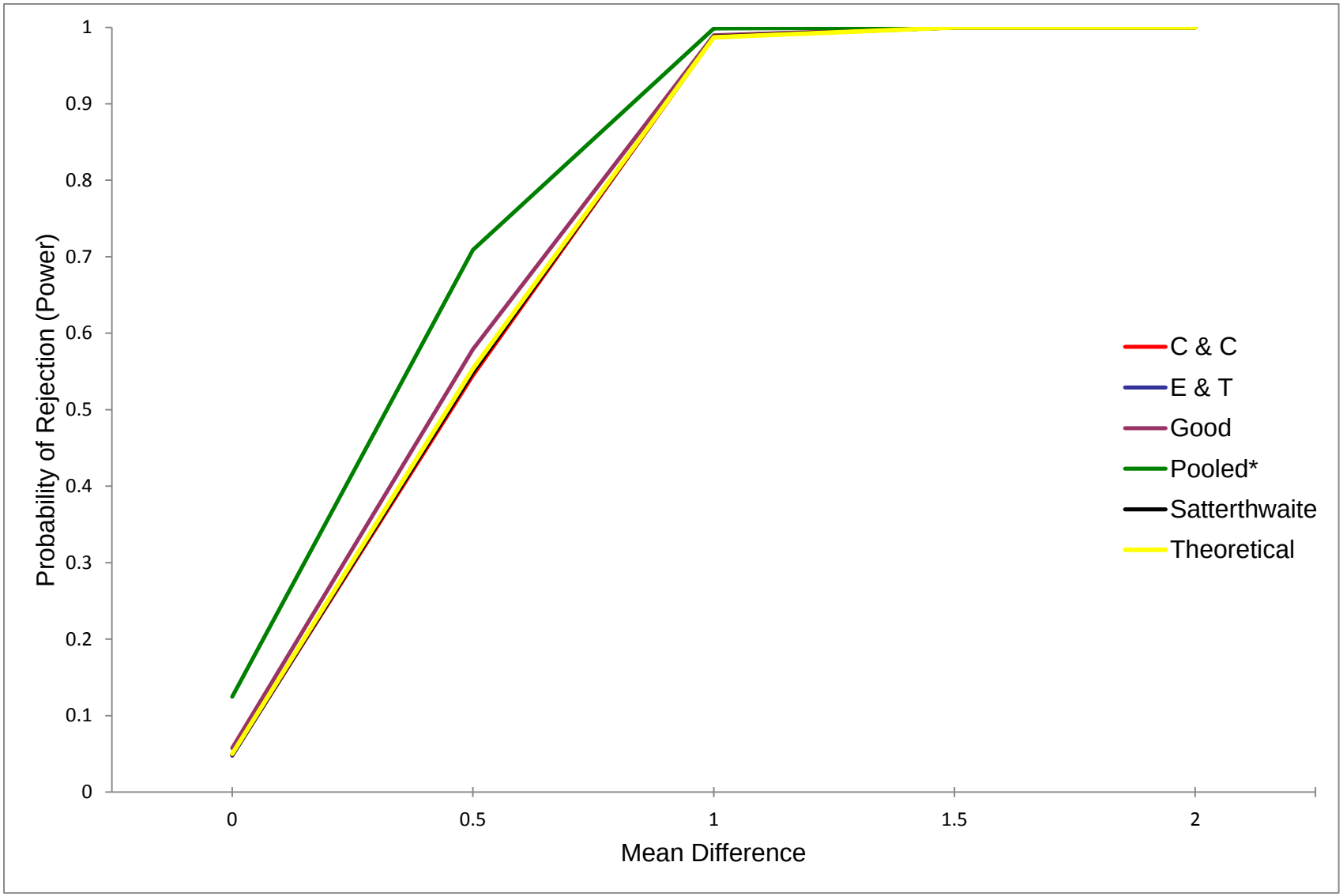

Figure C51. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

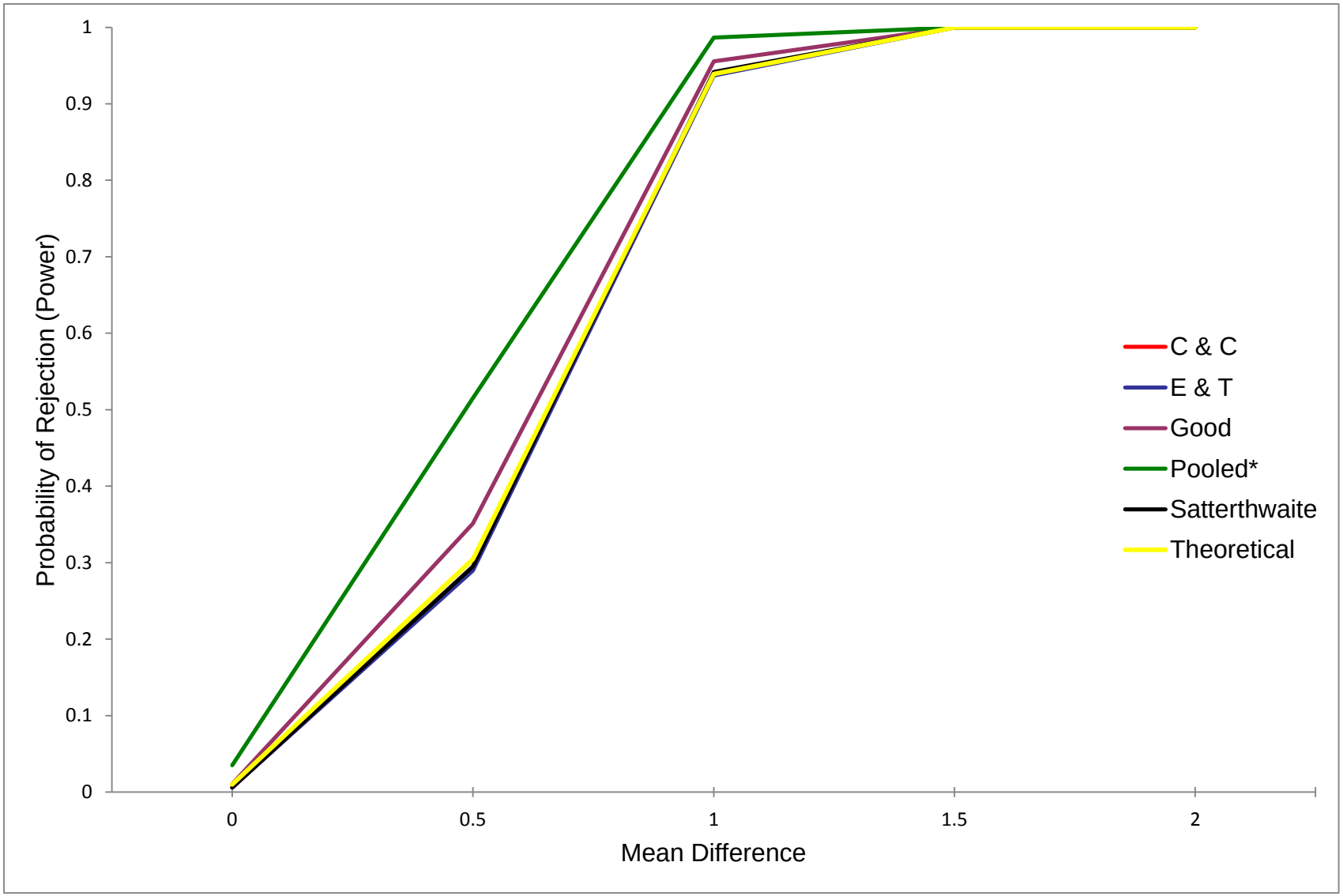

Figure C52. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 2, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

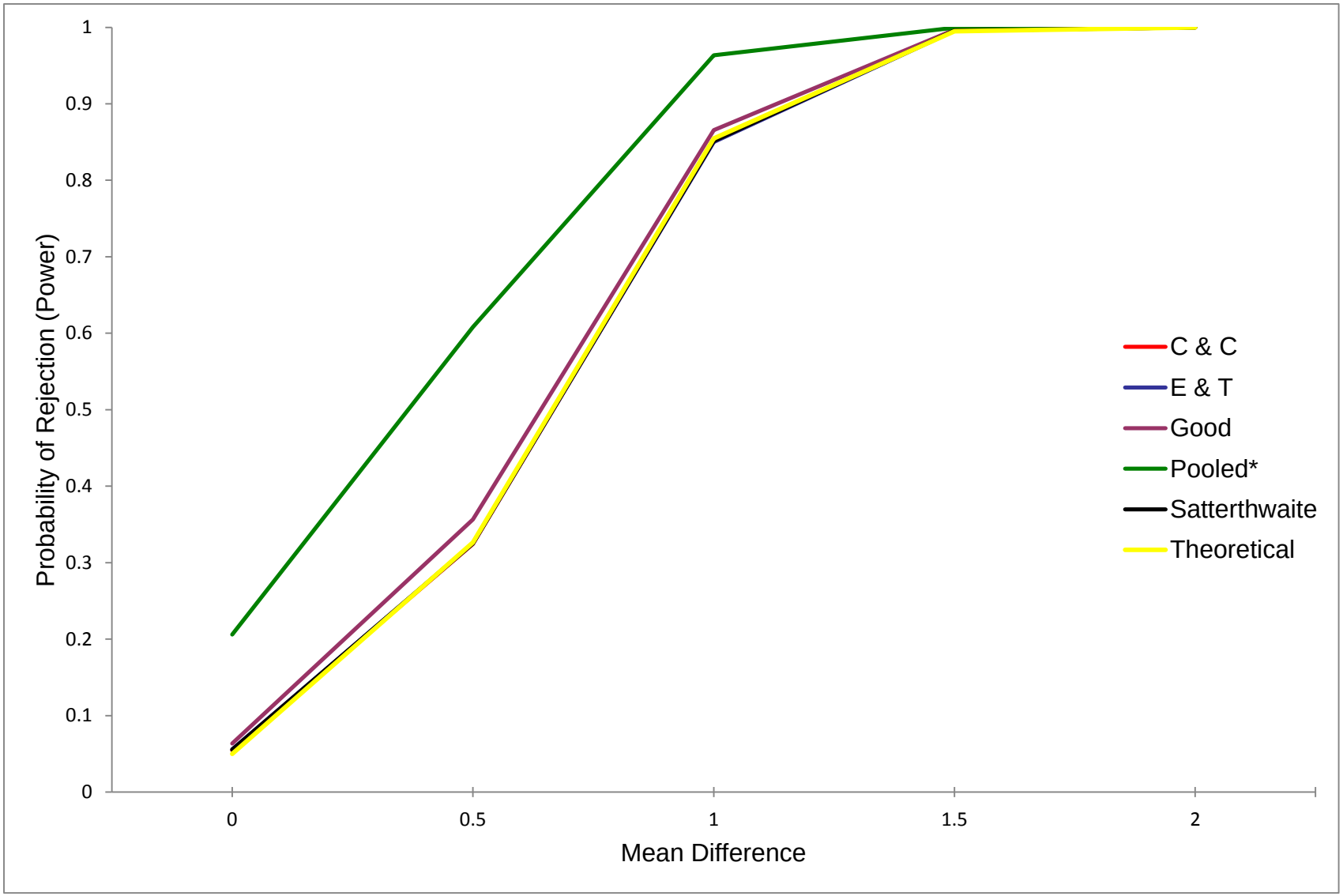

Figure C53. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

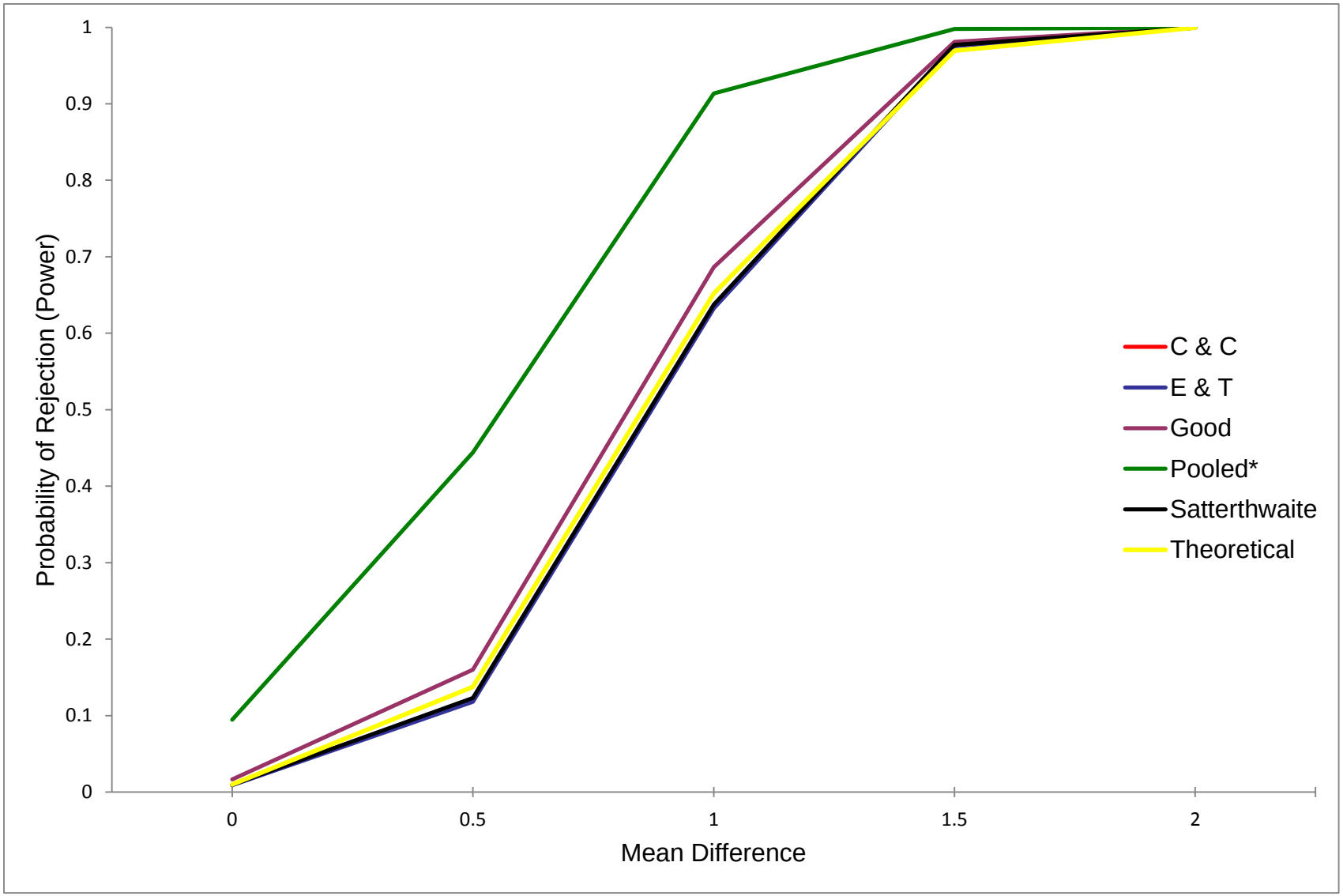

Figure C54. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 4, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

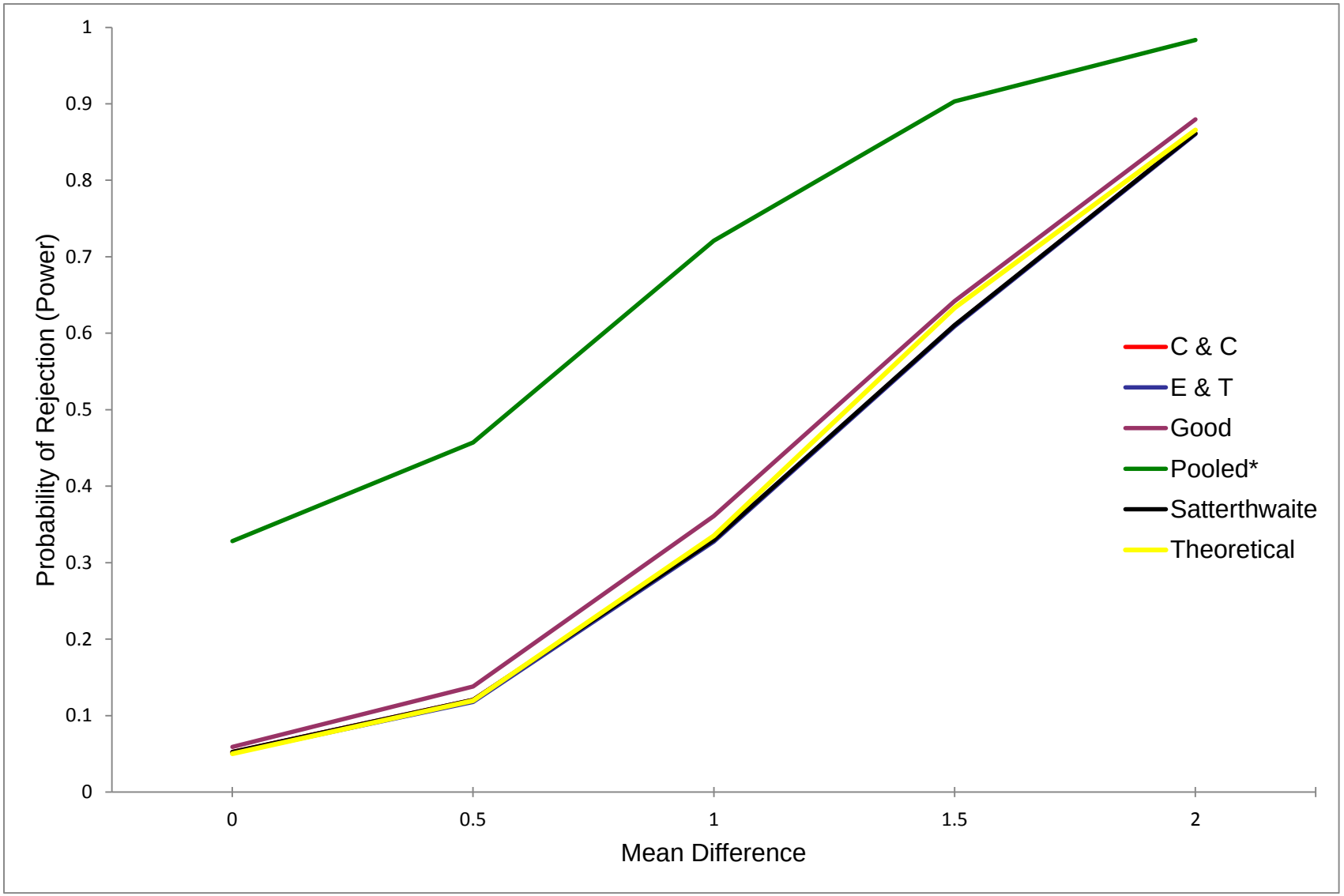

Figure C55. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .05. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

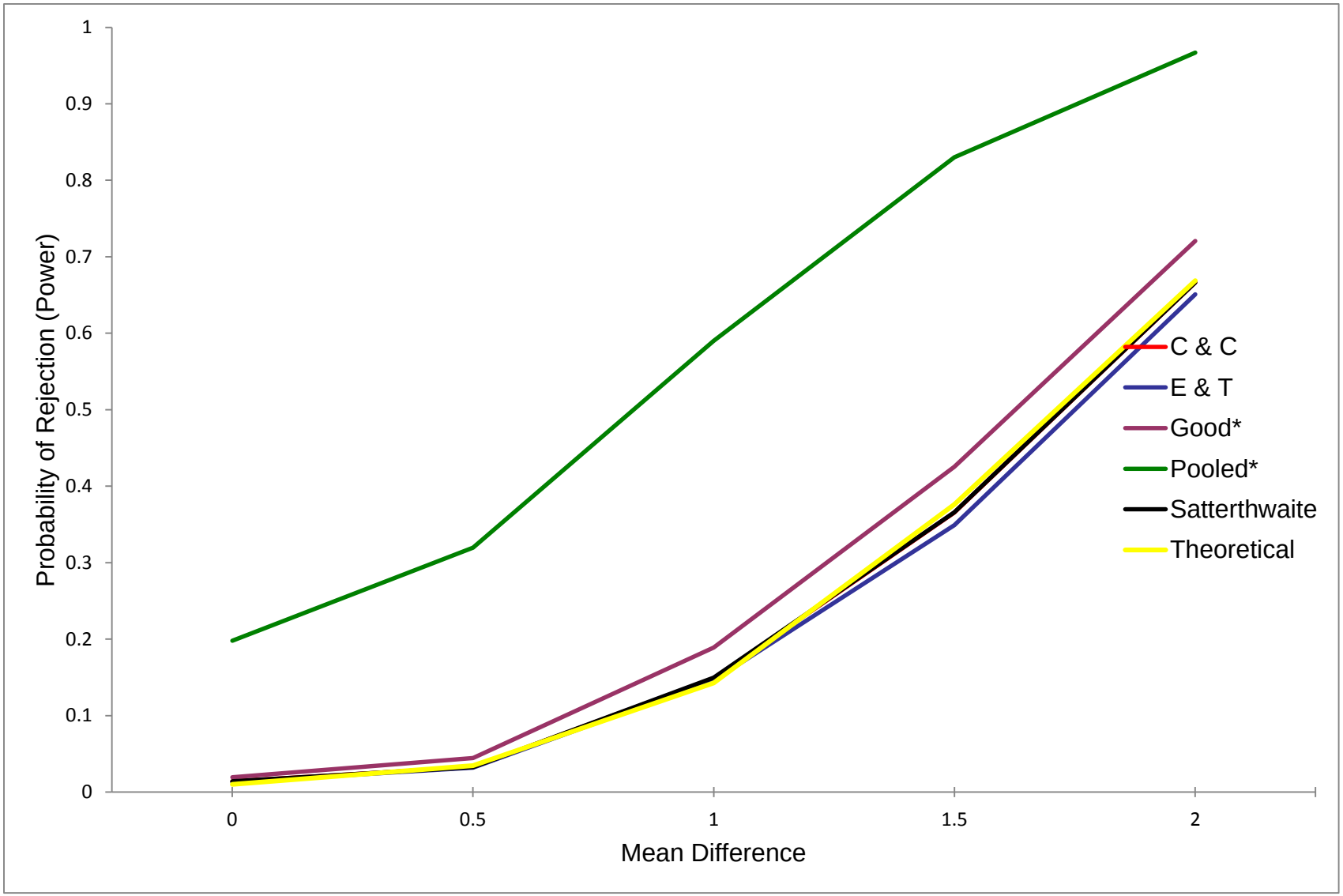

Figure C56. Power curves for unequal group sample sizes when $n_1 = 40$, $n_2 = 200$, and variance ratio (var_1/var_2) was 16, at a significance level (standard) of .01. C & C = Cochran and Cox; E & T = Efron and Tibshirani.
* Significant Type I error rate result ($p \leq .001$).

APPENDIX D: COMPARING THE POWER RESULTS OF THE MOST EXTREME POSITIVE AND NEGATIVE MEAN DIFFERENCES (I.E., μ_1-μ_2 = 2.0 VS. μ_1-μ_2 = -2.0)

Table D1

Comparing positive and negative mean differences power results[20] of each method based on 2,000 replications of the simulated experiment when $n_1 = 10$; $n_2 = 50$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .05

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
16	2.0	0.29	0.26	0.42	0.72	0.29
	-2.0	0.29	0.25	0.42	0.74	0.29

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Table D2

Comparing positive and negative mean differences power results[21] of each method based on 2,000 replications of the simulated experiment when $n_1 = 10$; $n_2 = 50$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .01

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	0.99	0.97	1.00	1.00	0.99
	-2.0	0.99	0.98	1.00	1.00	0.99
16	2.0	0.11	0.08	0.25	0.59	0.11
	-2.0	0.09	0.07	0.25	0.62	0.10

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

[20] Rounded to two decimal places
[21] Rounded to two decimal places

Table D3

Comparing positive and negative mean differences power results[22] of each method based on 2,000 replications of the simulated experiment when $n_1 = 25$; $n_2 = 125$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .05

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
16	2.0	0.64	0.63	0.70	0.93	0.64
	-2.0	0.67	0.66	0.72	0.93	0.67

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Table D4

Comparing positive and negative mean differences power results[23] of each method based on 2,000 replications of the simulated experiment when $n_1 = 25$; $n_2 = 125$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .01

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
16	2.0	0.38	0.36	0.48	0.87	0.38
	-2.0	0.40	0.37	0.50	0.88	0.40

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

[22] Rounded to two decimal places
[23] Rounded to two decimal places

Table D5

Comparing positive and negative mean differences power results[24] of each method based on 2,000 replications of the simulated experiment when $n_1 = 40$; $n_2 = 200$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .05

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
16	2.0	0.86	0.86	0.88	0.98	0.86
	-2.0	0.88	0.88	0.90	0.98	0.88

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

Table D6

Comparing positive and negative mean differences power results[25] of each method based on 2,000 replications of the simulated experiment when $n_1 = 40$; $n_2 = 200$; var_2 was fixed to 1; variance ratio = var_1/var_2; at a significance level (standard) of .01

Variance ratio	Mean difference	Methods				
		C&C	E&T	Good	Pooled	Satterthwaite
1/16	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
1	2.0	1.00	1.00	1.00	1.00	1.00
	-2.0	1.00	1.00	1.00	1.00	1.00
16	2.0	0.65	0.65	0.72	0.97	0.65
	-2.0	0.68	0.67	0.74	0.97	0.68

Note. C&C = Cochran and Cox; E&T = Efron and Tibshirani.

[24] Rounded to two decimal places
[25] Rounded to two decimal places

REFERENCES

Abdi, H. (2007). The Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 103–107). Thousand Oaks, CA: Sage.

Academic Technology Services, Statistical Consulting Group. (2012). How can I bootstrap estimates in SAS? Los Angeles: UCLA. Retrieved September 20, 2012 from http://www.ats.ucla.edu/stat/sas/faq/bootstrap.htm

Alba Fernández, V., Jiménez Gamero, M. D., & Muñoz García, J. (2008). A test for the two-sample problem based on empirical characteristics functions. *Computational Statistics and Data Analysis, 52*, 3730–3748.

Aspin, A. A., & Welch, B. L. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika, 36*, 290–296.

Beasley, W. H., & Rodgers, L. J. (2009). Resampling methods. In R. E. Millsap and A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 362–386). Los Angeles, CA: Sage.

Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen [A contribution to error estimation with few observations]. *Landwirtschaftliche Jahrbücher, 68*: 807–37.

Beran, R. (1986). Simulated power functions. *The Annals of Statistics, 14*, 151-173.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*, 49-64.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science*, *18*, 168–174.

Brunner, E., & Munzel, U. (2000). The non parametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal, 42*, 17–25.

Chernick, M. R., & LaBudde, R. A. (2011). *Bootstrap methods with applications to R.* Hoboken, NJ: John Wiley & Sons.

Cochran W. G., & Cox, G. M. (1950). *Experimental designs*. New York: John Wiley & Sons.

Davenport, J. M., & Webster, J. T. (1975). The Behrens-Fisher problem: An old solution revisited. *Metrika*, *22*, 47–54.

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews, 19*, 55–68.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*, 96–109.

Dudewicz, E. J., Ma, Y., Mai, E., & Su, H. (2007). Exact solutions to the Behrens-Fisher problem: Asymptotically optimal and finite simple efficient choice among. *Journal of Statistical Planning and Inference, 137,* 1584–1605.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Efron, B. (2001). The bootstrap and modern statistics. In A. E. Raftery, M. A. Tanner, and M. T. Wells (Eds.), *Statistics in the 21st century* (pp. 326–332). Boca Raton, FL: CRC Press.

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science 2004, Vol. 19(4),* 676–685.

Fan, X., Felsövályi, Á., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.

Fisher, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics, 6,* 391–398.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42,* 237-288.

Ghosh, M., & Kim, Y. -H. (2001). The Behrens-Fisher problem revisited: A Bayes-frequentist synthesis. *The Canadian Journal of Statistics, 29*, 5–17.

Good, P. I. (2005). *Resampling methods: A practical guide to data analysis* (3rd ed.). New York, NY: Birkhäuser Boston.

Good, P. I. (2013). *Introduction to statistics through resampling methods and R* (Second edition). Hoboken, NJ: John Wiley & Sons, Inc.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hayes, A. F. (2000). Randomization tests and the equality of variance assumption when comparing groups. *Animal Behaviour, 59*, 653–656.

Heiser, D. A. (2006). Statistical tests, tests of significance, and tests of a hypothesis using Excel. *Journal of Modern Applied Statistical Methods, 5*, 551–566.

Herrington, R. S. (2001). *Simulating statistical power curves with the bootstrap and robust estimation* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses (304715013).

Howell, D. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury Thompson Learning.

Hyslop, T., & Lupinacci, P. J. (2003). A nonparametric fitted test for the Behrens-Fisher problem. *Journal of Modern Applied Statistical Methods, 2*, 414–424.

Kasuya, E. (2001). Mann-Whitney *U* test when variances are unequal. *Animal Behaviour, 61*, 1247–1249.

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods, 2*, 288–309.

Kim, S. -H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics, 23*, 356–377.

Kohr, R. L., & Games, P. A. (1974, April). Procedures for testing $\mu_1 = \mu_2$ with unequal N's and variances. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Lee, A. F. S., & Gurland, J. (1975). Size and power of tests for equality of means of two populations with unequal variances. *Journal of the American Statistical Association, 70*, 933–941.

Levene, H (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford, CA: Stanford University Press.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics, 35*, 615–645.

Moore, D. S., & McCabe, G. P. (2004). *Introduction to the practice of statistics* (5th ed.). New York, NY: W. H. Freeman.

Neuhäuser, M., Lösch, C., & Jöckel, K. -H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis, 51*, 5055–5060.

Olejnik, S., & Luh, W. -M. (1994). Type I error rates, power, and sample sizes for two-stage solutions to the Behrens-Fisher problem when population distributions are non-normal. *Computational Statistics & Data Analysis, 17*, 409–420.

Pesarin, F. (1995). A new solution for the generalized Behrens-Fisher problem. *Statistica, 55*, 131–146.

Pesarin, F. (2001). *Multivariate permutation tests: With applications in biostatistics*. West Sussex, UK: John Wiley & Sons.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. West Sussex, UK: John Wiley & Sons.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*, 461–472.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association, 65*, 1501–1508.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Smith, H. F. (1936). The Problem of comparing the results of two experiments with unequal errors. *Journal of the Council of Science in Industrial Research, 9*, 211-212.

Stonehouse, J. M., & Forrester, G. J. (1998). Robustness of the *t* and *U* tests under combined assumption violations. *Journal of Applied Statistics, 25*, 63–74.

van Belle, G., Fisher L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A methodology for health sciences* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Wang, H., & Chow, S. -C. (2002). A practical approach for comparing means of two groups without equal variance assumption. *Statistics in Medicine, 21*, 3137–3151.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*, 356–362.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*, 28–35.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York, NY: John Wiley & Sons.